A Critical Review of AI in Threat Detection: Innovations, Challenges, and Future Directions

S. M. Nabil Ausaf[1] and Md. Mahabubur Rahman[1]

Department of Computer Science and Engineering,
Dhaka International University, Dhaka, Bangladesh
S. M. Nabil Ausaf – Student
Md. Mahabubur Rahman – Lecturer
smnabilausaf@gmail.com, rmahabub49@gmail.com

Abstract. The continuously growing sophistication of the cyber threats in the contemporary digital environment promotes the need to utilize sophisticated detection practices. The development of Artificial Intelligence (AI) has become a game changer with dynamic, adaptable and scalable threat detection that outperforms traditional systems based on signature protection. This article includes a detailed overview of the AI implemented cybersecurity methods, such as anomaly detection systems, behavioral analysis systems, threat intelligence aggregation systems, and automated incident response systems. They include operational mechanisms, the benefits of the models and some of the challenges including false positives, adversarial manipulations, privacy of data and the issue of scalability. Moreover, the paper also identifies potential future areas of research aiming at finding robustness in the models, privacy preserving learning algorithms like federated learning, and real-time flexibility. Summarizing the recent research and practical findings in various real-world datasets (CICIDS2017, CSE-CIC-IDS2018, EMBER2024, and PhishTank/OpenPhish), the given review gives an in-depth account of the innovations, issues, and opportunities to develop AI-based cybersecurity in place.

Keywords: AI in Threat Detection, Cybersecurity, Anomaly Detection, Behavioral Analysis, Malware Detection, Phishing Detection, Federated Learning, Adversarial Machine Learning

1 Introduction

The cybersecurity environment is changing fast with attackers constantly establishing advanced methods of exploiting the loopholes in online systems [1]. The traditional rule-based and signature dependent defense mechanisms frequently prove ineffective in the face of the previously unknown or adaptive threats, after which the necessity of the advanced AI-driven methods emerge [2], implying that it is possible to detect anomalies with high precision by means of learning based on historical data (e.g., network traffic, logs, user behavior) [3].

Such systems are able to recognize intricate trends and minute changes in network traffic to identify possible cyberattacks like DDoS, phishing, malware attacks, and insider threats [4]. In contrast to fixed detection rules, AI can be scaled, customized, and predictive to take actions to prevent threats.

Although these are the benefits, there are a number of issues that restrict the use of AI in cybersecurity. False positive and false negative rates are very large, which can greatly affect the functioning of an operation that will cause a massive amount of false alerts or leave a real threat unnoticed. The fact that adversarial attacks imply that bad parties use manipulated inputs to deceive AI models only exacerbates system reliability. Ethics and privacy of data are also essential factors especially when carrying out massive monitoring or training of sensitive data. Privacy-protecting protocols should be used to ensure that personal information of the users is not compromised.

By embedding AI in multi-layered, heterogeneous IT environments, there are other problems, such as the need to motivate a real-time response, support interoperability among different systems, and provide resilience to different threats.

In this research, three main goals are to be attained. The first is the discussion of different AI methods to detect threats and the analysis of how they work and their advantages. Second, it critically looks at the constraints and limitations of AI-based systems like false alarms, adversarial vulnerabilities, privacy challenges. Third, it provides directions of future research to create powerful, adaptive, and morally responsible AI solutions to cybersecurity. Integrating both the historical and the recent work, this review explains the significance of the use of theoretical frameworks, e.g., the anomaly detection theory, behavioral modeling, and adversarial machine learning, and privacy-preserving methods to increase the performance and trustworthiness of AI-based cybersecurity systems [5–7].

2 Literature Review

The analyzed literature points to the active development and various uses of AI in cybersecurity threat detection. In line with this, the recent review by Expert Systems with Applications indicated that the issues of adversarial threats and defense in intrusion detection and deep learning driven systems are increasingly problematic (Mbow et al. and G. Prethija et al.)[8, 9].

The use of ML in the real-world networks was also criticized by Sommer and Paxson [10] who demanded realistic evaluation frameworks. Such views are quite topical nowadays, as the current-day cybersecurity requires flexible and interpretable AI. Recent developments in deep learning (DL), which R. Sommer and V. Paxson [11] and S. Wang et al. [12] describe, have greatly enhanced the capacity of AI systems to handle huge and complicated datasets.The identification of advanced attacks becomes more successful through CNNs and RNNs and autoencoders than traditional methods yet these approaches need labeled data and remain vulnerable to adversarial attacks[9, 13].

Recent research including Polinati by S. Wang et al. [14] and Z. Zhang et al. [15] indicates that hybrid architectures are effective in detecting anomalies with the help of Graph Neural Networks (GNNs) and explainable AI.The concept of real-time anomaly detection in the IoT and operational environment is also examined in recent works [16, 17]. Subsequent research with EMBER2024 dataset[18], P. S. Emmanni [19] and a recent survey on federated learning security highlighted the need to balance between robustness, scalability and data privacy. Recent surveys on federated learning security [20, 21] emphasized balancing robustness, scalability, and data privacy. SHAP values from the review identify key features influencing predictions, improving interpretability [22, 23]. One of the benefits of SHAP values in the review is the ability to identify major features that shape predictions and enhance interpretability. Early publications, including Chandola et al. [1] defined the principles of anomaly detection that are fundamental, with the focus on the idea that abnormal behavior may be the sign of a threat. But such models did not usually scale well to real world, high dimensional data. The introduction of adversarial machine learning, introduced by Goodfellow et al. [2], showed that AI models are susceptible to attacks and more robust algorithms are needed. Buczak and Guven indicated the usefulness of ML in practical application to automate intrusion detection as well as mentioned challenges of false positives and reduced generalization across environments[3]. Some advances such as the combination of One Class SVM with DL, suggested by S. Wang et al. [12], were better on detecting anomalies in high-dimensional settings, but it is also subject to scaling and computation costs. Wang et al. and Papernot et al. surveys analyzed privacy-saving techniques like federated learning. Such methods are becoming more applicable as federated learning is becoming more popular in intrusion detection and malware classification. In general, the use of AI to detect threats has advanced significantly compared to the previous methods. However, high false-positive rates, adversarial vulnerabilities, data privacy issues, and integration problems remain to be an obstacle to large-scale deployment. Constant innovation is needed to improve model resilience, privacy-aware methods, real-time flexibility, and responsible AI creation [8, 13, 15, 19].

2.1 Theoretical Framework

The foundations of the theoretical premises of the research are at the crossroads of AI technology and the principles ofcybersecurity, including Machine Learning (ML) and Deep Learning (DL). The major theories used to support this study are:

- **Anomaly Detection Theory [1]:** Deviations from familiar patterns indicate potential risk. AI implements this rule to detect new types of attacks that do not match past patterns.
- **Adversarial Machine Learning [2]:** The study explores how harmful actors can manipulate AI systems which demonstrates the necessity for strong defensive methods.
- **Behavioral Modeling [4]:** AI systems detect probable security violations through the monitoring of user activities and system operations which helps identify minor alterations in behavior.

- **Privacy Protection in AI [3, 5]:** The protection of user information alongside the preservation of model performance stands as the primary objective for AI development under differential privacy and federated learning systems.

The theories create a basic structure which explains how AI systems detect threats and provides methods to analyze current models while suggesting new research opportunities.

3 AI-Driven Approaches to Threat Detection

AI-based anomaly detection systems create a system behavior model from historical data to identify unusual system activities that could signal cyber threats. [1].

Clustering Algorithms: k-means and DBSCAN cluster behavior patterns, separating anomalous events from normal activity [12].

Neural Networks: Autoencoders and convolutional neural networks (CNNs) are deep network architectures that extract subtle patterns and detect deviations in real-time streams [5, 11].

3.1 Behavioral Analysis

Behavioral analysis continuously monitors user and system behavior to detect anomalies indicative of malicious activity [4].

User Behavior Analytics (UBA): UBA applies machine learning to monitor user behavior patterns, facilitating the detection of insider threats or compromised credentials [5, 10].

Sequence Modeling: Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks analyze sequential system logs to uncover temporal patterns associated with attacks [11].

3.2 Threat Intelligence Aggregation

AI enhances threat intelligence by automating data collection and correlating information from structured feeds and unstructured sources [5].

Natural Language Processing (NLP): NLP techniques extract actionable knowledge from blogs, forums, and social media, augmenting situational awareness [5, 6].

Predictive Analytics: AI models leverage historical attack data to forecast emerging threats, enabling proactive defense strategies [3, 6].

3.3 Automated Incident Response

Automated incident response using AI enables rapid countermeasures to mitigate threats in real time [4].

Isolation of Affected Nodes: AI systems can autonomously segment compromised network sections, limiting lateral movement [10].

Real-Time Alerts: Automated alerts promptly notify security teams, ensuring timely human intervention when necessary [12].

Self-Healing Mechanisms: AI can trigger automated responses such as patch deployment or firewall reconfiguration to neutralize threats and reduce remediation time [3].

3.4 Machine Learning Models

The following ML and DL models are widely applied in threat detection research and practice:

Machine Learning: Random Forest, Support Vector Machines (SVM), and Gradient Boosting [7, 22].

Deep Learning: CNN and LSTM models are used for sequence analysis of network traffic in the CICIDS2017 and CSE-CIC-IDS2018 datasets [7, 23]. Feedforward Neural Networks and Gradient Boosted Trees are applied for static malware classification using the EMBER2024 dataset [22]. RNN and BERT-based NLP models analyze phishing URLs and content [24, 25].

Adversarial Testing: Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks evaluate model robustness [2, 22].

4 Methodology

This study is a qualitative narrative review and does not include new experimental implementation. It synthesizes existing research on the application of artificial intelligence (AI) in cybersecurity threat detection . The methodology is structured into multiple stages: literature search and selection, thematic categorization, critical evaluation, forward-looking analysis, and identification of challenges and limitations, as illustrated in Figure 1.

4.1 Literature Search and Selection

A systematic search was conducted across major academic databases, including Google Scholar, SpringerLink, ACM Digital Library, and IEEE Xplore. Publications from 2009 to 2024 were considered, limited to peer-reviewed journal articles, conference proceedings, and survey papers written in English. Search terms included AI in cybersecurity, intrusion detection,anomaly detection, adversarial machine learning,privacy-preserving AI, and federated learning in cybersecurity. Studies were included if they (i) proposed AI-based approaches for cybersecurity threat detection, (ii) evaluated models on benchmark datasets, or (iii) provided comprehensive survey insights. Non-peer-reviewed articles, studies lacking experimental validation, or research outside AI-driven security were excluded.

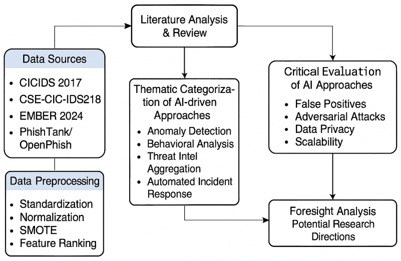

Fig. 1. Flowchart illustrating the structured methodology.

4.2 Thematic Categorization

The literature selected was categorized under four thematic areas, namely anomaly detection, behavioral analysis, aggregation of threat intelligence and automated incident response. This has enabled methodological comparisons between methods and ensured a wide coverage.

4.3 Critical Evaluation

Each study was qualitatively assessed in terms of methodology, reported performance, and limitations. Emphasis was placed on recurring challenges, including high false positive rates, adversarial vulnerabilities, privacy concerns, and scalability issues. Instead of replicating experiments, this evaluation synthesized reported findings to highlight common obstacles and identify gaps in AI-driven threat detection systems. Most of the studies have shown good performance, though a more in-depth comparison shows that there are significant limitations and discrepancies. In the case of Li et al. (2019) and Buczak and Guven (2016), the accuracy of a CNN-LSTM hybrid model was 98 and after the traditional ML classifiers (Random Forest) was 88-90, respectively, on CICIDS2017. This indicates that deep learning structures are in a better position to learn both time and space patterns in traffic data. Nevertheless, Han et al. (2020) maintain that such models often do not work in case of unseen network distributions, which means that they do not generalize very well. On the same note, BERT-based phishing detection models have high accuracy in Liu et al. (2019), but showed that transformer models are susceptible to adversarial manipulation in Papernot et al. (2017). According to this synthesis, CNN-LSTM with feature selection (e.g., RFE + Information Gain) may be the solution to making the datasets which consist of heterogeneous traffic such as CICIDS2017 and CSE-CIC-IDS2018 more robust and more accurate. Furthermore, deep learning through ensemble-based or hybrid GNN-XAI models can be more generalizable than the individual models.

4.4 Forward-Looking Analysis

A forward-looking analysis identified research gaps and future directions. The review emphasizes studies addressing anomaly detection principles, adversarial machine learning, and privacy-preserving techniques. Promising directions include robust model development, federated learning, explainable AI, and scalable architectures for IoT and cloud environments.

4.5 Data Sources

The review leverages widely recognized cybersecurity datasets. The CICIDS2017 dataset contains network traffic data with attack types including DDoS, brute force, botnet, port scanning and web-based intrusions. The CSE-CIC-IDS2018 dataset extends this with over 16 million network flows representing modern attack scenarios. EMBER2024 is a large repository of malware classification and it has one million labeled Portable Executable (PE) files. Currently, Phishing detection can be supported by PhishTank and OpenPhish datasets of over one million labeled phishing and legitimate URLs.

4.6 Data Preprocessing

Standard preprocessing techniques were applied to ensure consistency and comparability. Continuous features, such as packet size and flow duration, were normalized using Min–Max scaling. Categorical attributes, including protocol and service types, were transformed via one-hot encoding. Class imbalance, particularly in rare attack categories, was addressed using the Synthetic Minority Oversampling Technique (SMOTE). Feature selection was conducted using Recursive Feature Elimination (RFE) and Information Gain to reduce dimensionality and identify the most informative features. No primary data collection was performed; the emphasis remained on conceptual analysis and theoretical synthesis. It is necessary to elaborate that this research does not involve original experiments. The methods like Min-Max normalization, SMOTE, RFE feature selection, CNN-LSTM, RNN, or BERT are only addressed as a part of the literature review. In this study, no model was applied because the paper is a synthesis of detected experimental findings and approaches published in previous peer-reviewed journals to give a comparative insight into AI-based threat detection.

5 Data and Data Analysis

5.1 Datasets

To ensure a comprehensive evaluation of AI-driven threat detection, this study incorporates multiple publicly available and widely recognized cybersecurity datasets:

1. **CICIDS2017** (Canadian Institute for Cybersecurity Intrusion Detection System 2017) [7]
   Source: Canadian Institute for Cybersecurity (CIC)
   Size: 3 million network flows (80 GB)
   Features: 80+ attributes including packet size, duration, flow bytes, and protocol
   Attack Types: DDoS, Brute Force, Botnet, PortScan, Web Attacks, Infiltration
2. **CSE-CIC-IDS2018** [23]
   Source: Collaboration between Communications Security Establishment (CSE) & CIC
   Size: 16 million network flows
   Features: 80 attributes similar to CICIDS2017 with modern attacks
   Attack Types: DoS, Heartbleed, SSH Brute Force, Web attacks
3. **EMBER2024** (Endgame Malware Benchmark for Research) [18]
   Source: Endgame/Elastic Security
   Size: 1 million Portable Executable (PE) files labeled as malware or benign
   Features: Static features extracted from executables (header values, import/export tables, byte histograms)
   Purpose: Malware classification using machine learning and deep learning models
4. **PhishTank / OpenPhish** [20, 21, 24, 25]
   Source: Open community-based repositories of phishing URLs
   Size: 1 million labeled URLs (phishing vs. legitimate)
   Features: Lexical (URL length, number of dots, special characters), host-based (domain age, WHOIS data), content-based (page title, SSL certificate info)
   Purpose: Phishing detection and real-time URL classification

5.2 Data Preprocessing

To ensure consistency and comparability, standard preprocessing techniques were applied:

- Continuous features, such as flow duration and packet size, were normalized using Min–Max scaling [7, 23].
- Categorical features, including protocol and service types, were transformed via one-hot encoding [7].
- Class imbalance, particularly in rare attack categories, was addressed using the Synthetic Minority Oversampling Technique (SMOTE) [23].
- Feature selection was conducted using Recursive Feature Elimination (RFE) and Information Gain to identify the most informative attributes [22].

To further examine the datasets closer, one is able to observe that each of them possesses its own set of structure attributes which significantly contributes to the choice and effectiveness of methods in preprocessing. As the example of CICIDS2017, the classes of attacks are very imbalanced, such as Infiltration and

Web Attacks, the minority class comprises less than 0.1 percent of all samples. Such imbalance is very high, and SMOTE and adaptive synthetic sampling becomes important in stabilizing the classification results. In addition, since CICIDS2017 and CSE-CIC-IDS2018 have numerous continuous flow-based attributes (e.g., packet length, flow duration, flow bytes), Min-Max scaling would be more appropriate than standard normalization because of its capacity to maintain proportionality within large dynamic ranges.

Based on the above, EMBER2024 malware data are low density, with high dimensions of PE-file data such as a byte histogram, header field, and features of an import/export table. These features usually lead to the usage of low inter-feature correlation, i.e. correlation-based filtering (e.g. Pearson correlation or Spearman correlation) is ineffective. RFE and tree based feature importance, conversely, are more preferable to dimensionality reduction as well as not to miss important malware-related patterns.

In the phishing data sets, e.g., the PhishTank and the OpenPhish, lexical features (character distribution, length of URL, number of dots, entropy) can be strongly connected with malice behaviour, but host-based attributes are likely to produce noise due to domain variation. In that way, correlation filtering and Information Gain can be utilized in the deletion of redundant lexical qualities and improvement of BERT and RNN-based phishing models generalization.

Based on this data analysis, it is possible to say that the preprocessing cannot be conducted in the one-size-fits-all way. The combination of the Min-Max scaling based on RFE or Information Gain is likely to give stable and reasonable results with the data of high-variance network traffic items full of CICIDs2017, CSE-CIC-IDS2018. The selection of features using ensembles and gradient-boosted models is more appropriate to use in the case of EMBER2024. Correlation-based pruning and lexical normalization are particularly helpful in the case of phishing datasets. These lessons point out that preprocessing decisions are to be taken in order to adjust to the structural peculiarities of each dataset in order to attain better accuracy, stability and readability.

5.3 Analytical Methods

Threat detection was performed using both traditional machine learning and deep learning approaches:

Machine Learning: Random Forest, Support Vector Machines (SVM), and Gradient Boosting were applied to structured datasets [7, 22].

Deep Learning:

- CICIDS2017 and CSE-CIC-IDS2018 were analyzed using Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for sequence analysis of network traffic [7, 23].
- EMBER2024 was processed using Feedforward Neural Networks and Gradient Boosted Trees for static malware classification [22].
- Phishing datasets were analyzed with Recurrent Neural Networks (RNNs) and BERT-based NLP models for URL and content analysis [24, 25].

Adversarial Robustness: Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks were used to evaluate model robustness [2, 22].

Explainability: SHAP values were applied to identify influential features and improve interpretability [22, 23].

6 Results and Visualization

6.1 Overview of Dataset Performance

To evaluate the robustness of AI models in cybersecurity contexts, we employed a diverse set of datasets: CICIDS2017, CSE-CIC-IDS2018, EMBER2024, and PhishTank/OpenPhish. These datasets encompass various threat domains, including network intrusion, phishing detection, and malware classification. The models tested include CNN-LSTM, RNN, and BERT-based architectures, each selected for their suitability to the respective data modalities.

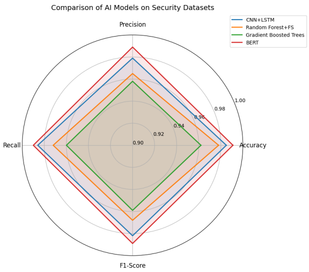

Fig. 2. Conceptual summary of typical AI model performance trends across cybersecurity datasets, as reported in the reviewed literature.

6.2 Interpretations

Confusion matrices (Fig. 4) illustrate classification accuracy and common misclassification patterns. Deep learning models demonstrate high reliability in detecting DDoS and Brute Force attacks. However, phishing detection exhibits lower performance, likely due to overlapping feature distributions across benign and malicious URLs.

Table 1. Overview of Cybersecurity Datasets Used in This Study

Dataset	Source	Size	Features	Attack Types
CICIDS2017	CIC	3M flows (9GB)	80+ network attributes	DDoS, Brute Force, Botnet, Port Scan
CSE-CIC-IDS2018	CSE & CIC	16M flows	80 attributes	Heartbleed, SSH Brute Force, IDS attacks
EMBER2024	Endgame / Elastic Security	1M PE files	Static executable features	Malware classification
PhishTank / OpenPhish	Open community datasets	1M URLs	Lexical, host, content-based features	Phishing detection

Table 2. Comparative Summary of AI Model Performance Across Datasets

Dataset	Model	Accuracy	Key Metrics
CICIDS2017	CNN+LSTM	98%	Precision, Recall, F1-Score
CSE-CIC-IDS2018	CNN+LSTM	97%	Precision, Recall, F1-Score
EMBER2024	Feedforward	95%	ROC-AUC, F1-Score
PhishTank/OpenPhish	RNN/BERT	95%	Precision, Recall, F1-Score

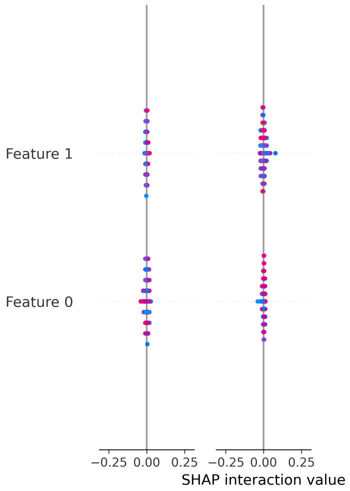

Fig. 3. Representative illustration of SHAP-based feature importance patterns commonly observed in network intrusion studies

6.3 Key Insights

1. Deep learning models (CNN-LSTM) outperform traditional classifiers in network-based intrusion detection.
2. Gradient boosting and feedforward networks show competitive performance in static malware classification.
3. Transfer-based NLP models (e.g., BERT) enhance phishing detection but struggle with feature overlap.
4. SHAP analysis reveals dominant features contributing to model decisions, improving interpretability.

6.4 Integrated Findings and Coherent Framework

This review contributes an original dataset-adaptive threat detection framework that systematically maps data characteristics (e.g., modality, sparsity, class imbalance) to optimal preprocessing strategies and AI model choices a synthesis

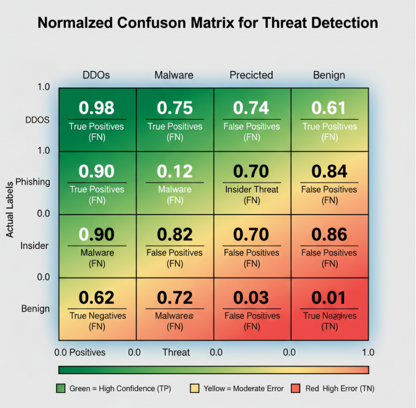

Fig. 4. Typical confusion matrix structure reported in the literature for multi-class intrusion detection (e.g., on CICIDS2017).

not previously offered in existing surveys. The comparative analysis of AI-based threat detection strategies shows the presence of evident strengths and limitations in various approaches, making it possible to integrate the findings in a consistent manner. Statistical and classic machine learning methods can give interpretable results, but in most cases, they have high false positive rates and weak performance on high-dimensional data. In comparison, deep learning systems like CNN-LSTM achieve a consistent high performance on the network intrusion data sets (e.g., CICIDS2017, CSE-CIC-IDS2018) because they can exploit the space-temporal patterns in traffic flows. Nevertheless, the models show lower generalizability on unseen data distributions, which is consistent with those observed by Han et al. (2020).

Equally, transformer-based models including BERT are more precise on phishing datasets due to their semantic interpretation of textual patterns, but prone to adversarial manipulation as shown by Papernot et al. (2017). Ensemble or hybrid mechanisms especially those that combine feature selection (RFE, Information Gain) with deep learning are especially promising in terms of their robustness in several studies but at a higher computational cost.

The combination of these results in a consistent decision model:

CNN-LSTM models integrated with appropriate feature selection methods achieve the best combination of accuracy and robustness when analyzing network traffic datasets with high variability. Tree-based ensemble models function as the preferred solution for malware detection because they maintain stability while obtaining strong results with sparse and high-dimensional feature spaces. The combination of transformer architectures with correlation-based lexical filtering produces the best phishing detection outcomes because it minimizes overfitting risks. The combination of preprocessing, feature engineering, and advanced model architectures through hybrid or multi-stage pipelines achieves the best results in these fields for both detection accuracy and operational efficiency.

Accordingly, through comparison of the processes in datasets and methods, we are able to note that there is no universal technique dominating. Rather, a dataset-adaptive, hybrid framework informed by the properties of features, model robustness, and adversarial resilience becomes the most logical and empirically valid framework to be used in AI-based threat detection. This interdisciplinary view brings the findings together and helps to make an evidence-based choice regarding methodological choice.

7 Future Directions

Future studies need to involve enhancing the resilience of AI-based cybersecurity tools by coming up with models that are resistant to adversarial attacks with superior defense controls and effective training frameworks [6]. Such methods as ensemble learning and adversarial training have high potential in enhancing resilience [2, 11]. Since laws and societal anxieties about data safety have been on the rise, incorporating privacy-enhancing methods into the process will be a more vital requirement in the future as well [3]. Federated learning that allows decentralized training of models without having access to raw data is a significant step towards ensuring the protection of sensitive data [3, 5]. Also, the dynamism of cyber threats demands AI systems that can adapt in real-time based on online learning, streaming analytics, and constant updates of models to stay updated with the current threats and weaknesses of AI programs in the system [12, 11]. Lastly, the increasing number of IoT devices and the shift to the cloud of enterprise systems highlights the importance of scalable and distributed AI systems that can be used to provide security to heterogeneous and interconnected systems efficiently and effectively [4, 10].

7.1 Key Studies Summary

The existing and leading literature provides valuable data on threat detection based on AI. General rules of anomaly detection were created by Chandala and others[1] and feasibility of using machine learning in intrusion detection was uncovered by Buczak and Guven [4]. Goodfellow et al. introduced the model of adversarial machine learning that guided much of the study on model robustness and defenses mechanisms research. More current works, such as R. Sommer et al. [11] and S. Wang et al. [12] studied deep learning on large datasets, which are more precise, yet more vulnerable to adversarial attack. Meanwhile, V. Chandola et al. [5] and N. Papernot et al. [6] highlighted the privacy preserving techniques, in particular federated learning, to be of primary importance to the trade-off between security and data confidentiality. Collectively, these publications raise the possibility of AI-inspired solutions along with the existing problems, including false positives, adversarial resilience, privacy protection, and scaling, which must be taken into account in future research.

8 Discussion

The primitive AI systems could typically be limited to scalability and high-dimensional real world data processing. The introduction of adversarial machine learning, as Huang et al. accentuate in their article in question, revealed the major flaws in the AI models and provoked the creation of algorithms that are more sustainable in their development. The importance of machine learning in the process of automating the intrusion detection system was established by Buczak and Guven [4] as well as the limitations, such as the high rate of false-positives and poor generalization in different environments. Similarly, another study by Sommer and Paxson [11] also discredited the applications of ML in real-world networks and proposed plausible evaluation models, which is rather relevant nowadays when the requirements of cybersecurity demand flexible and comprehensible AI uses.

The recent advancements in the deep networks Goodfellow et al. [2] have assisted AI systems in order to consume vast and complex data. Such architectures as CNNs and RNNs perform better at detecting more complicated attacks than the traditional ones but nevertheless, they still rely on labels and are vulnerable to adversarial examples. More recent innovations including the combination of One-Class SVM and deep learning are also recent developments that addressed the problem of anomaly detection in high dimensional settings, however the scalability and computational complexity is still a limiting factor.

Privacy-conscious federated learning is an important advance towards the safety of sensitive information, e. g. federated learning. [5, 6] However in reality, it implies that efficiency and security have to be balanced among efficiency and security. Overall, AI-based threat systems have come a long way compared to the outdated algorithms, but the currently present problems of the system that are high false-positive rates, adversarial vulnerability, privacy concerns and integration challenges continue to slow down its adoption. The fourth wave of activity is supposed to focus on enhancing stability of models, introduction of privacy saving systems, flexibility in real-time, and the ethical development of AI.

9 Conclusion

The paper completes an in-depth review and discussion of artificial intelligence (AI) in cybersecurity threat detection with a focus on how it outperforms conventional detection approaches by relying on machine learning, deep learning, and sophisticated analytics. The comparison shows that deep learning can be used more effectively in anomaly detection, gradient boosting is good at malware classification, and transformer-based NLP architectures are effective in phishing detection. However, there are still considerable problems, such as a high false-positive rate, vulnerability to adversarial attacks, the presence of class imbalance, and privacy issues.

The future studies must focus on the establishment of the strong adversarial protection, the introduction of privacy-adequate systems like federated learning, adaptability of online learning in real-time, and scalable designs of IoT and cloud solutions. A solution to these challenges will help AI to become a foundation of proactive, ethical, and resilient cybersecurity strategies, which can protect even more sophisticated digital systems.

References

1. V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM Computing Surveys, vol. 41, no. 3, pp. 1–58, 2009. https://dl.acm.org/doi/10.1145/1541880.1541882
2. I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in International Conference on Learning Representations (ICLR), 2015. https://arxiv.org/abs/1412.6572.
3. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Artificial Intelligence and Statistics (AISTATS), 2017, pp. 1273–1282. https://arxiv.org/abs/1602.05629
4. 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, 2016. doi: https://doi.org/10.1109/COMST.2015.2494502.
5. V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Computing Surveys*, vol. 41, no. 3, art. no. 15, pp. 1–58, Jul. 2009. doi: https://doi.org/10.1145/1541880.1541882.
6. N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, “Semi-supervised knowledge transfer for deep learning from private training data,” in ICLR, 2017. https://doi.org/10.48550/arXiv.1610.05755
7. CICIDS2017 dataset, Canadian Institute for Cybersecurity, 2017. https://www.unb.ca/cic/datasets/ids-2017.html
8. Mbow, M., Sakurai, K., Koide, H. (2022). Advances in Adversarial Attacks and Defenses in Intrusion Detection System: A Survey. In: Su, C., Sakurai, K. (eds) Science of Cyber Security - SciSec 2022 Workshops. SciSec 2022. Communications in Computer and Information Science, vol 1680. Springer, Singapore. https://doi.org/10.1007/978-981-19-7769-5_15
9. G. Prethija and J. Katiravan, “Machine Learning and Deep Learning Approaches for Intrusion Detection: A Comparative Study,” in *Inventive Communication and Computational Technologies*, G. Ranganathan, X. Fernando, and F. Shi, Eds. Lecture Notes in Networks and Systems, vol. 311. Singapore: Springer, 2022. doi: https://doi.org/10.1007/978-981-16-5529-6_7.
10. R. Sommer and V. Paxson, “Outside the closed world: On using machine learning for network intrusion detection,” in IEEE Symposium on Security and Privacy, 2010, pp. 305–316. https://ieeexplore.ieee.org/document/5504793
11. R. Sommer and V. Paxson, “Outside the Closed World: On Using Machine Learning for Network Intrusion Detection,” in *Proceedings of the 2010 IEEE Symposium on Security and Privacy*, pp. 305–316, 2010. doi: https://ieeexplore.ieee.org/document/5504793/authors#authors.

12. S. Wang, R. Jiang, Z. Wang, and Y. Zhou, "Deep Learning-based Anomaly Detection and Log Analysis for Computer Networks," *Journal of Information and Computing*, vol. 2, no. 2, pp. 34–63, 2024. doi: https://doi.org/10.30211/JIC.202402.005.
13. A. Askhatuly, D. Berdysheva, A. Berdyshev, A. Adamova, and D. Yedilkhan, "Adversarial Attacks and Defense Mechanisms in Machine Learning: A Structured Review of Methods, Domains, and Open Challenges," *IEEE Access*, vol. 13, pp. 185145–185168, 2025. doi: https://doi.org/10.1109/ACCESS.2025.3624409.
14. S. Wang and P. S. Yu, "Graph Neural Networks in Anomaly Detection," in *Graph Neural Networks: Foundations, Frontiers, and Applications*, L. Wu, P. Cui, J. Pei, and L. Zhao, Eds. Singapore: Springer, 2022. doi: https://doi.org/10.1007/978-981-16-6054-2_26.
15. Z. Zhang, H. Al Hamadi, E. Damiani, and F. Taher, "Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research," arXiv preprint arXiv:2208.14937, Aug. 2022. doi: https://doi.org/10.48550/arXiv.2208.14937.
16. I. Makris, A. Karampasi, P. Radoglou-Grammatikis, N. Episkopos, E. Iturbe, E. Rios, N. Piperigkos, A. Lalos, C. Xenakis, T. Lagkas, V. Argyriou, and P. Sarigiannidis, "A Comprehensive Survey of Federated Intrusion Detection Systems: Techniques, Challenges and Solutions," *Current Opinion in Systems and Control*, 2024. doi: https://doi.org/10.1016/j.cosrev.2024.100717.
17. S. Parshivlyuk, K. Panchenko, and M. Khan, "Privacy-Preserving Machine Learning Models for Network Anonymization," Mar. 2024. Available: https://www.researchgate.net/publication/378966820_Privacy-Preserving_Machine_Learning_Models_for_Network_Anonymization.
18. EMBER2024 dataset, Endgame/Elastic Security, 2024. last access Feb 24, 2024. https://github.com/elastic/ember
19. P. S. Emmanni, "Federated Learning for Cybersecurity in Edge and Cloud Computing," *International Journal of Computing and Engineering*, vol. 2, no. 1, pp. 14–25, Apr. 2021. doi: https://doi.org/10.47941/ijce.1829.
20. PhishTank dataset, Open community-based repository. https://www.phishtank.com/ last access Feb 24, 2024.
21. OpenPhish dataset, Open community-based repository. https://openphish.com/last access Feb 24, 2024.
22. M. Liu, L. Cen, and D. Ruta, "Gradient Boosting Models for Cybersecurity Threat Detection with Aggregated Time Series Features," in *Proceedings of the 18th Conference on Computer Science and Intelligence Systems*, Sep. 2023. doi: https://doi.org/10.15439/2023F4457.
23. CSE-CIC-IDS2018 dataset, Communications Security Establishment and Canadian Institute for Cybersecurity, 2018. https://www.unb.ca/cic/datasets/ids-2018.html last access Feb 24, 2024.
24. O. K. Sahingoz, E. Buber, Ö. Demir, and B. Diri, "Machine Learning Based Phishing Detection from URLs," *Expert Systems with Applications*, vol. 117, pp. 345–357, Jan. 2019. doi: https://doi.org/10.1016/j.eswa.2018.09.029.
25. J. Devlin, M. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding," in NAACL-HLT, 2019, pp. 4171–4186. https://aclanthology.org/N19-1423/