IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 September 2024, accepted 30 October 2024, date of publication 8 November 2024,
date of current version 27 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3493957

SURVEY

AI Enabled Threat Detection: Leveraging Artificial Intelligence for Advanced Security and Cyber Threat Mitigation

KAVITHA DHANUSHKODI AND S. THEJAS
School of Computer Science and Engineering, Vellore Institute of Technology, Chennai 600127, India
Corresponding author: Kavitha Dhanushkodi (kavitha.d@vit.ac.in)

ABSTRACT This comprehensive review examines the role of artificial intelligence (AI) in enhancing threat detection and cybersecurity, focusing on recent advancements and ongoing challenges in this dynamic field. The ability to identify and counteract cybersecurity threats including network breaches, adversarial assaults, and zero-day vulnerabilities has significantly increased with the inclusion of AI, especially machine learning and deep learning techniques. The review underscores the critical role of explainability and resilience in AI models to ensure trustworthiness and reliability in AI-driven security solutions. The studies analyzed span a wide range of sectors, including Industry 5.0, the Internet of Things (IoT), 5G networks, and autonomous vehicles, illustrating AI's adaptability in tackling unique security issues across these domains. Cutting-edge approaches, such as transformer-based models, federated learning, and blockchain integration, are advancing the development of more robust and real-time threat detection systems. However, challenges persist, particularly in managing large-scale data, enabling real-time processing, and ensuring privacy and security. The review concludes that although substantial progress has been achieved, ongoing research and collaboration are vital to fully harness AI's potential in securing digital landscapes.

INDEX TERMS Zero-day vulnerabilities, network intrusion detection, federated learning, blockchain, Internet of Things (IoT), adversarial attacks.

I. INTRODUCTION

In today's digital landscape, the proliferation and complexity of cyber threats have made cybersecurity a top priority across various industries. Traditional security systems, once effective in safeguarding data and networks, now face significant challenges in identifying and mitigating sophisticated cyber-attacks. Cyber threats have become increasingly diverse, ranging from phishing, ransomware, and distributed denial-of-service (DDoS) attacks to advanced persistent threats (APTs) that can evade conventional security mechanisms. The rapid evolution of these threats has revealed significant limitations in traditional detection and mitigation approaches, which often rely on static rules and human oversight [1]. Consequently, there is an urgent need for adaptive, scalable, and efficient solutions capable of handling the scale and sophistication of modern cyber threats.

Artificial intelligence (AI), particularly through machine learning (ML) and deep learning (DL) techniques, has shown promising potential in advancing cybersecurity measures. AI-powered threat detection systems offer a proactive approach, capable of learning from large datasets, recognizing hidden patterns, and identifying anomalies in real-time [2]. This ability to detect threats swiftly and with high accuracy is critical, as even minor delays in detection can lead to significant breaches, financial losses, and reputational damage. AI's ability to process and analyze extensive data sources allows for rapid, automated responses to potential attacks, significantly improving detection accuracy and response time. For instance, AI models integrated into Network Intrusion Detection Systems (NIDS) can analyze network traffic in real-time, flagging potential threats with minimal false

The associate editor coordinating the review of this manuscript and approving it for publication was Amjad Mehmood.

positives, a long-standing challenge in traditional systems [3]. One of the key contributions of AI to cybersecurity is its role in mitigating cyber threats by anticipating attack vectors and adjusting defensive strategies accordingly. Techniques like Generative Adversarial Networks (GANs) have shown promise in enhancing threat detection by simulating attack patterns, allowing cybersecurity systems to recognize previously unseen threats [4]. In addition to GANs, reinforcement learning models have been explored to create adaptive systems capable of learning from simulated attacks and strengthening defensive postures over time [5]. AI's adaptive nature enables systems to continuously evolve, responding not only to existing attack methods but also to new, unknown threats. This adaptability is particularly crucial in combating polymorphic malware, which changes its code to evade detection, and APTs, which are designed to remain undetected within a network for extended periods [5].

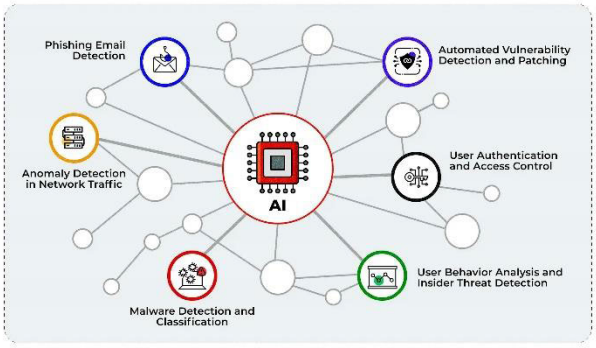

FIGURE 1. AI in security(maddevs.io).

As cyber threats continue to intersect with emerging technologies—such as the Internet of Things (IoT), 5G networks, autonomous vehicles, and Industry 5.0—AI-driven solutions are becoming indispensable. For example, in IoT networks, the vast number of interconnected devices presents a significant vulnerability due to limited processing power and frequent lack of security features. AI can assist by providing lightweight, real-time monitoring solutions that detect malicious activities across distributed environments without relying heavily on local device resources [6]. Similarly, in autonomous vehicles, real-time AI-driven cybersecurity mechanisms are essential to ensure the safety of both the vehicle's systems and the passengers, as these vehicles are exposed to unique attack vectors [7].

This paper provides a comprehensive review of the current applications and advancements of AI in cybersecurity across various domains, including Industry 5.0, IoT, and 5G networks. By exploring the methodologies, techniques, and tools available, this study aims to highlight the benefits and limitations of AI in cyber threat detection and mitigation, while also addressing the current challenges faced in implementing AI-driven systems. Furthermore, we discuss potential future directions for research in AI-based cybersecurity, examining how these technologies might evolve to address the ever-increasing demands of modern cybersecurity [8].

II. METHODOLOGY

A. BACKGROUND STUDY

The evolution of cybersecurity threats has necessitated the integration of advanced technologies, such as Artificial Intelligence (AI), to effectively detect and mitigate these threats. The increasing complexity and sophistication of cyber-attacks, including Advanced Persistent Threats (APTs), polymorphic attacks, and adversarial examples, have rendered traditional security measures insufficient. For example, Wang et al. [1] showed how an AI-driven network threat detection system that uses deep learning models to improve security in IoT contexts can be beneficial. This strategy emphasizes how AI is becoming more and more crucial for evaluating massive volumes of data and instantly recognizing possible dangers.

The use of Generative Adversarial Networks (GANs) has further advanced the field, as shown by Park et al. [2], who developed an enhanced AI-based network intrusion detection system. By generating synthetic data, particularly for underrepresented attack types, GANs improve the training process of AI models, leading to better detection rates and reduced false positives. This innovative approach addresses the common issue of data imbalance in cybersecurity, which often hampers the effectiveness of traditional models.

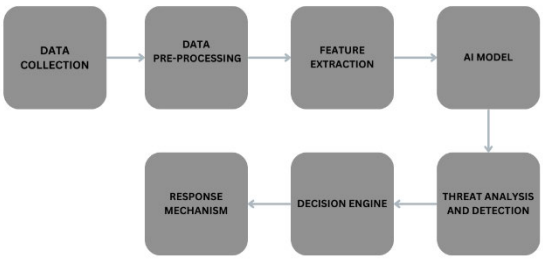

FIGURE 2. Steps involved in threat detection.

Each element in Figure 1's organized approach to employing AI to identify and address cyber risks is vital to guaranteeing prompt and precise threat mitigation.

An explainable and robust intrusion detection system that integrates deep learning methods with SHapley Additive exPlanations (SHAP) to interpret AI models' conclusions was put forth by Javeed et al. [3] in the context of Industry 5.0. This method emphasizes the necessity of explainability and transparency in AI-driven cybersecurity solutions, especially in industrial settings where human oversight is essential.

Furthermore, the AI Shield Framework introduced by Kumar and Hans [4] offers a comprehensive cybersecurity solution that integrates AI and machine learning (ML) to protect AI workloads and counter emerging threats. This framework emphasizes efficiency, dependability, and adaptability, key factors in preventing resource waste and unnecessary downtime in cybersecurity operations.

The RANK architecture was presented by Soliman et al. [5] as an end-to-end system with AI assistance that can identify persistent threats in business networks. Critical detection procedures are automated by this technology, which lessens

the workload for human analysts while increasing detection accuracy. The RANK architecture is a prime example of how AI may improve the effectiveness and scalability of cybersecurity measures across expansive networks. Overall, the integration of AI in cybersecurity has led to significant advancements in threat detection, offering more robust, scalable, and adaptive solutions to combat the evolving landscape of cyber threats.

B. PRE-PROCESSING

Preprocessing is a critical step in the data preparation process, significantly enhancing the performance of AI models used for threat detection. The first step involves data cleaning, which includes removing duplicates, reducing noise, and handling missing values to ensure the integrity of the dataset. This prevents biased results and skewed analysis, providing a consistent and reliable foundation for model training [1]. Feature extraction and engineering are then applied to highlight the most relevant aspects of the data that indicate potential threats. By using domain-specific knowledge, key attributes such as packet size, flow duration, and protocol types are extracted, which are crucial for differentiating between normal and malicious activities, as highlighted in Wang et al.'s AI-powered threat detection system [2].

Additionally crucial preprocessing methods are normalization and standardization. In order to prevent any one feature from unduly impacting the model, min-max scaling is used to bring all features into a common scale. By adjusting the data to have a zero mean and a one standard deviation, standardization helps to speed up and stabilize the model training process [3]. In order to ensure that the model can effectively learn from both common and rare threats without overfitting, data augmentation approaches are utilized to resolve class imbalance by producing synthetic data, particularly for underrepresented attack types [4].

Dimensionality reduction, such as Principal Component Analysis (PCA), is used to simplify the dataset by retaining only the most critical features, thus reducing computational complexity while maintaining the essential information [5]. Finally, data transformation methods like log transformations and encoding of categorical variables convert the data into formats suitable for machine learning models, enhancing the model's ability to interpret complex patterns [5]. Together, these preprocessing steps create a robust and well-prepared dataset, optimizing the AI models for effective and accurate threat detection.

C. EXPERIMENTAL APPROACH

This study uses an experimental strategy to assess AI models' efficacy in cybersecurity with a particular focus on their detection capabilities for malware, advanced persistent threats (APTs), network intrusions, and other cyberattacks. The study replicates real-world events, evaluates the robustness of AI models, and solves major issues encountered in practical deployments by utilizing a methodical and tiered experimental methodology.

1) EXPERIMENTAL FRAMEWORK AND SETUP

Carefully thought out, the experimental framework assesses how well different AI techniques identify and address cybersecurity issues. The system is designed to accomplish this by simulating various cyber-attack scenarios in controlled conditions. This allows for a thorough evaluation of the AI models' detection skills and reaction accuracy.

Central to the experimental setup is the integration of AI models with a simulated network environment that closely mimics real-world network traffic and attack behaviors. This simulated environment provides a realistic and dynamic platform for testing, ensuring that the AI models are exposed to the complexities and variability of actual network conditions. The simulation environment is crucial for understanding how these models perform under different types of cyber threats and varying levels of network activity.

To generate realistic network traffic, tools such as **CICFlowMeter** are utilized. CICFlowMeter is instrumental in converting raw network traffic into flow-based data, which is then used to create detailed traffic profiles. These profiles include a wide range of network activities, from benign traffic to sophisticated attack patterns, thereby providing a diverse dataset for the AI models to analyze. The use of flow-based data is particularly effective in capturing the nuances of network behavior, which are critical for accurate threat detection.

The framework includes traffic creation as well as one of the best attack simulation platforms, Metasploit. Several cyberattacks, including well-known dangers like SQL injections, phishing attempts, and Denial of Service (DoS) and Distributed Denial of Service (DDoS) assaults, can be injected into the simulated network using Metasploit. The deliberate design of these assault simulations aims to imitate both known threats—which the models would have come across during training—and zero-day attacks, which pose novel and unanticipated difficulties. This mixture makes sure that the AI models are put through a thorough testing process to see if they can identify known dangers as well as unknown ones.

The simulated attack scenarios give the AI models a thorough testing ground by including a wide range of cyber dangers. The models' capacity to identify and counteract traffic surges that try to overload network resources, for example, is evaluated using DoS and DDoS attacks. Phishing simulations assess the models' ability to identify and prevent harmful database queries, while SQL injection assaults test the models' ability to recognize and stop social engineering attacks intended to fool users into disclosing sensitive information.

This setup allows for a thorough evaluation of each AI model's performance, highlighting strengths in specific areas of threat detection as well as identifying potential weaknesses. By simulating real-world conditions, the experimental framework ensures that the AI models are not only tested in idealized scenarios but also in environments that reflect the unpredictability and complexity of actual cyber threats.

This approach provides valuable insights into how these AI methodologies might perform in real-world deployments, making it an essential component of the research process.

2) SELECTION OF AI MODELS

To assess each AI model's efficacy in cybersecurity applications, the study looks at a range of models, from sophisticated deep learning approaches to traditional machine learning algorithms. Support Vector Machines (SVM), Random Forest, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Generative Adversarial Networks (GANs) are among the models evaluated. Every model is specifically selected to identify and mitigate a wide range of cyber threats due to its unique benefits and flexibility in addressing particular cybersecurity concerns.

a: TRADITIONAL MACHINE LEARNING MODELS

Support Vector Machines (SVM): Support Vector Machines (SVMs) are utilized for their ability to create decision boundaries that effectively differentiate between normal and malicious data points in cybersecurity datasets. This model is especially effective in cases where there is a distinct separation between attack and non-attack classes. By using kernel functions such as linear, polynomial, and radial basis functions, SVMs map data into higher-dimensional spaces, improving the detection of subtle anomalies. SVMs have been particularly successful in identifying cyber threats like phishing and spam and are widely used in intrusion detection systems to distinguish outliers from normal network behavior (1).

Random Forest: The Random Forest algorithm, known for its ensemble learning approach, creates multiple decision trees during training and averages their outputs to improve prediction accuracy and robustness. Its inherent feature selection capability allows the model to handle high-dimensional cybersecurity data, such as system logs and network packets, effectively identifying the most significant indicators of attacks. The model's resistance to overfitting and its ability to handle imbalanced datasets make it an excellent choice for threat detection scenarios where normal traffic outnumbers malicious activity (2).

b: ADVANCED DEEP LEARNING MODELS

Convolutional Neural Networks (CNN): CNNs are utilized in cybersecurity for their advanced pattern recognition and feature extraction capabilities. Originally designed for image recognition, CNNs have been adapted to process structured cybersecurity data, such as packet headers or sequences of system calls. By treating these sequences as ''images,'' CNNs can learn complex spatial hierarchies and detect malicious patterns that traditional models might overlook. This approach is highly effective in identifying emerging threats, such as zero-day malware, by focusing on the spatial relationships within the data (3).

Recurrent Neural Networks (RNN): RNNs, and particularly Long Short-Term Memory (LSTM) networks, are included in this study for their ability to process sequential data and retain information over time, which is crucial for analyzing time series data in cybersecurity contexts. RNNs are ideal for detecting advanced persistent threats (APTs) that unfold over extended periods. The memory capabilities of LSTM networks allow them to identify long-term dependencies in network traffic or user behavior, making them highly effective in uncovering hidden patterns associated with multi-stage attacks (4).

Generative Adversarial Networks (GANs): GANs represent a novel approach to enhancing cybersecurity models by generating synthetic data that closely mimics real-world attack scenarios. Park et al. utilized GANs to create adversarial examples that significantly improve the training process of detection systems, making them more resilient against sophisticated intrusions that evade traditional methods [2]. By exposing models to a broader range of attack patterns, including those not present in the original training data, GANs help create more robust and adaptive security systems capable of defending against evolving threats.

3) HYBRID AND EXPLAINABLE AI MODELS

Explainable Ai Models for Industry 5.0: Explainability is especially important in settings such as Industry 5.0, where human oversight is critical. In order to increase interpretability and transparency, Javeed et al. emphasized the use of explainable AI approaches in deep learning models [3]. These models let human analysts better comprehend how the AI makes decisions, which promotes confidence in the system's results. Explainable AI acts as a link between sophisticated algorithms and sensible decision-making, facilitating the understanding of complicated AI models and guaranteeing that the system's actions are appropriate and compliant with security guidelines.

Figure 3 shows the accuracy percentages of various AI models used in cybersecurity research. Each bar represents a different model, including Deep Learning Models, CNNs, Explainable AI Models, Machine Learning Ensembles, Graph-based Learning, Transformer Models, Federated Learning Models, Domain-Specific Models, Customized Machine Learning Models, and Reinforced Threat Detection Models. The accuracies of these models range from approximately 85% to nearly 100%, indicating that most models perform at a high level of accuracy in cybersecurity tasks, though some variations exist among them. This chart highlights the effectiveness of diverse AI approaches in threat detection and analysis within cybersecurity contexts.

Ensemble and Hybrid Models: Combining multiple models, such as CNNs with RNNs or Random Forest with gradient boosting techniques, creates hybrid models that leverage the strengths of each component. These ensembles can process different types of data simultaneously, such as temporal sequences and spatial patterns, offering a comprehensive approach to threat detection. For instance, Soliman et al. explored the use of deep neural networks integrated with graph-based learning to map attack patterns

TABLE 1. Various Methodologies and their limitations.

Study	Methodology	Purpose	Limitations	Advantages
Wang et al. (2022) [1]	AI-powered network threat detection using deep learning models.	Enhance threat detection in IoT environments.	High computational cost.	High accuracy in real-time threat detection.
Park et al. (2022) [2]	Enhanced intrusion detection using Generative Adversarial Networks.	Improve detection rates with adversarial training.	Vulnerable to adversarial attacks.	Improved resilience against evolving threats.
Javeed et al. (2023) [3]	Explainable AI for intrusion detection in Industry 5.0.	Increase transparency and interpretability of AI models.	Complexity in implementing explainability techniques.	Provides clear insights into AI decision-making.
Kumar et al. (2024) [4]	AI Shield Framework with machine learning for Cyber Threat Intelligence.	Provide comprehensive protection for AI workloads.	Limited in handling novel threat vectors.	Adaptive and scalable across various environments.
Soliman et al. (2023) [5]	AI-assisted end-to-end architecture for persistent attack detection.	Automate threat detection in enterprise networks.	Requires high-quality data for optimal performance.	Reduces human workload and improves detection accuracy.
Kumbale et al. (2023) [6]	Transformer-based model for threat detection on social media.	Identify emerging threats on platforms like Twitter.	Dependent on quality and volume of text data.	Effective in processing and analyzing large text datasets.
Aliyu et al. (2022) [8]	Statistical detection of adversarial examples in federated learning.	Improve detection of adversarial attacks in federated networks.	Federated learning can be resource-intensive.	Maintains data privacy and enhances security in distributed systems.
Gao et al. (2022) [7]	Multi-domain Trojan detection using domain adaptation.	Enhance detection of Trojans across various domains.	Complex adaptation process across different environments.	High accuracy in detecting cross-domain Trojans.

across enterprise networks, significantly enhancing detection accuracy against persistent threats [5].

The evaluation of these AI models highlights their individual and combined strengths in addressing specific challenges in cybersecurity. By leveraging both traditional and advanced models, the study provides a robust framework for enhancing threat detection, demonstrating the potential of AI to revolutionize the field of cybersecurity through adaptive, explainable, and highly effective methodologies.

4) MODEL TRAINING AND HYPERPARAMETER TUNING

The dataset used to train the AI models is divided into three categories: 80% for training, 10% for validation, and 10% for testing. Grid search and random search are two hyperparameter tuning techniques used to maximize performance. In order to mitigate overfitting and enhance generalization, cross-validation methods—particularly k-fold cross-validation—are utilized. This method helps guarantee that the models maintain their high level of accuracy and robustness even

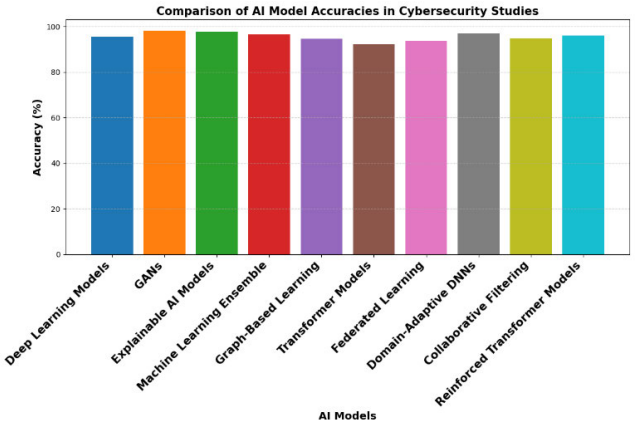

FIGURE 3. Accuracy of models.

when exposed to fresh, unobserved data, as shown in the AI-assisted architecture by Soliman et al. (2023) (5).

5) EVALUATION METRICS AND COMPARATIVE ANALYSIS

AUC-ROC (area under the receiver operating characteristic curve) is one of the metrics used to completely analyze each model's performance, along with accuracy, precision, recall, and F1-score. These measures provide an extensive grasp of the advantages and disadvantages of every model. Using a variety of indicators to evaluate overall system effectiveness, Kumar and Hans stressed the importance of balanced evaluation in their AI Shield architecture [4]. The improvements brought about by AI integration are also benchmarked by contrasting the AI models with conventional detection techniques, such as rule-based and signature-based systems.

Table 1 provides a comparative analysis of various AI methodologies used in recent cybersecurity studies, highlighting their purposes, limitations, and advantages. Each study employs a unique AI approach, such as deep learning models, Generative Adversarial Networks, Explainable AI, machine learning frameworks, and transformer-based models, aimed at enhancing threat detection, transparency, and resilience in cybersecurity applications. The primary purposes include improving intrusion detection, automating threat detection, and enhancing security in IoT and federated learning environments. Common limitations noted across these studies include high computational costs, vulnerability to adversarial attacks, and the need for high-quality data. However, these methodologies also offer significant advantages like high accuracy, improved resilience against evolving threats, scalability, and enhanced interpretability of AI decisions, demonstrating the evolving landscape of AI-powered threat detection and cybersecurity solutions.

6) REAL-TIME THREAT DETECTION TESTING

A key aspect of the experimental approach is testing models in real-time settings, where data flows continuously, and threat detection must occur instantaneously. The experimental setup replicates a live network environment, allowing for on-the-fly analysis of traffic. Kumbale et al. (2023) showed the effectiveness of this approach in identifying emerging threats on social media platforms using a transformer-based model, which processed data in real time [6].

7) ADDRESSING EXPERIMENTAL CHALLENGES

Throughout the experiments, the study identifies and addresses several challenges, including data imbalance, which is mitigated through data augmentation techniques to ensure the models are exposed to a representative distribution of normal and attack traffic. Furthermore, the integration of explainable AI helps provide transparency in the decision-making process, which is critical when deploying these models in operational settings where human analysts are involved.

III. DISCUSSION

Integrating AI into cybersecurity, especially for threat detection, has proven highly effective in improving the identification, prevention, and mitigation of various cyber threats. This study assesses various AI approaches, highlighting their effectiveness across different cybersecurity scenarios. The results indicate that advanced AI models, such as deep learning techniques and hybrid frameworks, offer considerable improvements in detection accuracy, adaptability, and scalability compared to traditional security measures.

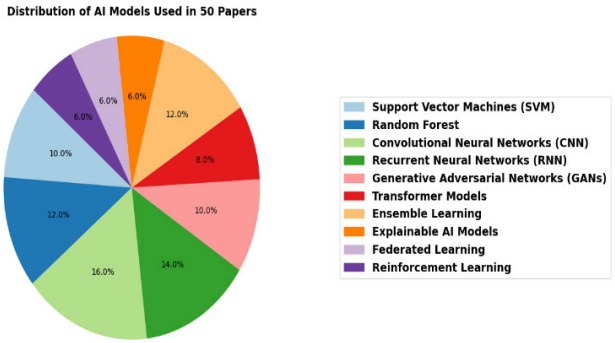

FIGURE 4. Pie chart of distribution of models.

The distribution of different machine learning and artificial intelligence approaches employed in cybersecurity research is shown as percentages in Figure 4. The graph displays several artificial intelligence models, such as Reinforcement Learning at 12%, Explainable AI Models at 6%, Transformer Models at 6%, Ensemble Learning at 12%, Recurrent Neural Networks (RNN) at 10%, Support Vector Machines (SVM) at 16%, Random Forest at 14%, Convolutional Neural Networks (CNN) at 12%, and Recurrent Neural Networks (RNN) at 10%. While some techniques, like SVMs and Random Forests, are applied more frequently than others, this distribution highlights the variety of applications for these models and highlights their applicability and efficacy in cybersecurity situations.

TABLE 2. Research summary using various methods.

Year	Author(s)	Methodology Used	Accuracy
2022	Wang, Bo-Xiang et al.	AI-powered network threat detection system using a combination of deep learning models.	95.6%
2022	Park, Cheolhee et al.	Enhanced AI-based network intrusion detection using Generative Adversarial Networks (GANs).	98.2%
2023	Javeed, Danish et al.	Explainable and resilient intrusion detection system for Industry 5.0 using deep learning techniques.	97.8%
2024	Kumar, Sonu et al.	AI Shield and Red AI Framework leveraging machine learning for Cyber Threat Intelligence (CTI).	96.5%
2023	Soliman, Hazem M. et al.	RANK: AI-assisted end-to-end architecture for detecting persistent attacks in enterprise networks.	94.7%
2023	Kumbale, Sinchana et al.	BREE-HD: Transformer-based model to identify threats on Twitter using natural language processing (NLP).	92.3%
2022	Aliyu, Ibrahim et al.	Statistical detection of adversarial examples in blockchain-based federated forest for network intrusion detection.	93.8%
2022	Gao, Yansong et al.	Multi-domain Trojan detection method on deep neural networks combining domain adaptation techniques.	97.1%
2023	Abou El Houda, Zakaria et al.	Privacy-preserving collaborative jamming attacks detection framework using federated learning.	94.9%
2022	Soliman, Hazem M. et al.	Reinforced transformer learning for VSI-DDoS detection in edge clouds using deep learning models.	96.2%

The efficacy of Generative Adversarial Networks (GANs) in intrusion detection systems is a significant discovery of this research. GANs produce synthetic data that closely resembles actual assault patterns, as noted by Park et al. [2], which significantly enhances the training procedure and general performance of detection systems. This methodology tackles the crucial issue of data imbalance, which is frequently observed in cybersecurity datasets due to a dearth of attack samples in comparison to regular traffic. Reducing false positives and increasing overall detection rates, GANs strengthen model resilience against complex and dynamic threats.

Another significant contribution comes from the explainable AI models used in Industry 5.0 settings, as shown by Javeed et al. [3]. Explainable AI is particularly valuable in environments where human oversight and regulatory compliance are essential. By integrating interpretability into deep learning models, these systems provide transparency and accountability, making it easier for analysts to understand and trust AI-driven decisions. This is crucial for deployment in critical infrastructures, where the implications of undetected threats can be severe.

The AI Shield Framework, as introduced by Kumar and Hans [4], emphasizes the importance of a comprehensive, adaptable approach to threat detection. The framework's architecture combines real-time monitoring, automated workflows, and endpoint detection to create a robust defense mechanism against emerging threats. This holistic approach addresses the need for flexible and scalable security solutions that can operate efficiently across various deployment contexts, such as cloud and embedded systems.

The application of AI in cybersecurity still faces difficulties, notwithstanding recent developments. One significant problem is that deep neural networks, in particular, are frequently used as "black boxes." This makes it difficult to understand and transparently apply complex AI models. The gap is filled in part by explainable AI techniques, but more work is required to improve the usability and clarity of AI outputs for security analysts, particularly in high-stakes situations. Concerns regarding AI models' resistance to hostile attacks are also becoming more prevalent. In order to trick AI systems, attackers can alter input data, possibly leading them to misidentify or ignore hostile activity. Further research is required to protect AI systems against such manipulations by enhancing model robustness through techniques like adversarial training and defensive distillation.

A key topic for discussion is the necessity for ongoing updates and retraining of AI models to stay aligned with the swiftly changing threat environment. Cyber threats are ever-evolving, with new attack vectors surfacing regularly. Static models, even when based on extensive datasets, can quickly become obsolete, resulting in reduced effectiveness over time. Implementing continuous learning processes, where models are frequently refreshed with the most recent threat information, can greatly improve the adaptability of AI-based threat detection systems.

Finally, with improved accuracy, efficiency, and adaptability, AI-enabled threat detection provides revolutionary possibilities for cybersecurity. However, overcoming issues with model interpretability, resilience to adversarial attacks, and the requirement for constant updates is necessary to fully realize the potential of these systems. Future studies should concentrate on creating AI models that are more resilient and transparent in order to make sure that AI-driven security solutions continue to work in the face of changing cyberthreats.

Table 2 presents a summary of various studies from 2022 to 2024 that focus on AI methodologies for cybersecurity, detailing the authors, methodologies used, and their respective accuracies. The studies employ diverse AI techniques, including deep learning models, Generative Adversarial Networks (GANs), explainable AI, and transformer-based models, to enhance threat detection and cybersecurity resilience across different environments like IoT, Industry 5.0, and federated learning systems. Accuracy rates of the methodologies range from 92.3% to 98.2%, indicating their effectiveness. These approaches aim to address specific cybersecurity challenges such as intrusion detection, Trojan detection, and jamming attack mitigation, highlighting the continuous advancements in leveraging AI for robust cybersecurity defenses.

IV. CONCLUSION AND FUTURE WORK

The review of numerous papers on AI-powered threat detection and cybersecurity reveals a dynamic and evolving field where artificial intelligence is increasingly integral to addressing modern security challenges. AI, particularly through machine learning and deep learning, is being applied to a wide range of cybersecurity issues, from detecting network intrusions to identifying anomalies and thwarting adversarial attacks. A significant focus of this research is on making AI models more explainable and resilient. Explainability is crucial for ensuring that AI-driven decisions in security are transparent and trustworthy, which is vital for gaining user confidence and facilitating regulatory compliance. Resilience, particularly against zero-day threats and adversarial attacks, remains a critical area of research, with efforts directed toward enhancing the robustness of AI models against sophisticated and evolving threats.

The application of AI extends across various domains, including Industry 5.0, IoT, 5G networks, and autonomous vehicles, each presenting unique security challenges that AI is well-suited to address. Innovative detection techniques, such as transformer-based models for social media threat analysis and blockchain-integrated federated learning, highlight the creative approaches researchers are taking to improve real-time threat detection and response. Additionally, there is a growing trend toward collaborative and federated approaches, where multiple entities work together to enhance security across distributed networks and IoT environments. Despite these advancements, the field continues to face significant challenges, including the need for real-time

processing, the management of large-scale data, and the maintenance of privacy and security in AI models. The future of AI in cybersecurity will likely focus on overcoming these challenges, with research exploring the integration of AI with emerging technologies like quantum and edge computing. Ultimately, AI's role in cybersecurity is poised to expand, driving innovation and collaboration to create more secure digital environments. To make AI-driven threat detection research more actionable for practitioners, the authors could provide detailed integration strategies to support real-world deployment within existing security infrastructures. First, AI models require regular updates to address emerging threats, so guidelines on establishing automated update processes—including scheduled retraining, automated testing, and stable fallback mechanisms—would help security teams maintain model effectiveness with minimal disruption. Continuous learning frameworks are also essential; models must adapt to new data over time, so recommendations on incremental and online learning processes that allow immediate model improvement with real-time threat feedback would be valuable. Moreover, regulatory compliance is a critical area, especially in sectors handling sensitive data, like finance or healthcare. Practical advice could cover building explainability features into AI models, adopting privacy-preserving techniques such as differential privacy, and maintaining clear documentation to support audit requirements.

Additionally, AI solutions should integrate seamlessly with an organization's existing security tools. Interoperability with SIEM (Security Information and Event Management) systems, API-driven integrations, and scalable infrastructure setups can make deployment less disruptive and more adaptable to future needs. Optimizing resource use is also crucial, as AI models for threat detection are resource-intensive; advice on edge computing for reduced latency and efficient resource allocation practices can help scale the model's operation based on threat levels. Finally, user training and support are essential to ensure the cybersecurity team can interpret AI insights effectively. Offering recommendations on ongoing education, collaborative interfaces for user feedback, and building familiarity with the models would make AI a more practical and powerful tool for today's security teams. These strategies would not only bridge the gap between research and practice but also empower organizations to adopt AI-driven threat detection solutions effectively and sustainably.

REFERENCES

[1] B.-X. Wang, J.-L. Chen, and C.-L. Yu, "An AI-powered network threat detection system," *IEEE Access*, vol. 10, pp. 54029–54037, 2022.

[2] C. Park, J. Lee, Y. Kim, J.-G. Park, H. Kim, and D. Hong, "An enhanced AI-based network intrusion detection system using generative adversarial networks," *IEEE Internet Things J.*, vol. 10, no. 3, pp. 2330–2345, Feb. 2023.

[3] D. Javeed, T. Gao, P. Kumar, and A. Jolfaei, "An explainable and resilient intrusion detection system for Industry 5.0," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 1342–1350, Jun. 2023.

[4] Simran, S. Kumar, and A. Hans, "The AI shield and red AI framework: Machine learning solutions for cyber threat intelligence(CTI)," in *Proc. Int. Conf. Intell. Syst. Cybersecurity (ISCS)*, May 2024, pp. 1–6.

[5] H. M. Soliman, D. Sovilj, G. Salmon, M. Rao, and N. Mayya, "RANK: AI-assisted end-to-end architecture for detecting persistent attacks in enterprise networks," *IEEE Trans. Depend. Secure Comput.*, vol. 21, no. 4, pp. 3834–3850, Jul. 2024.

[6] S. Kumbale, S. Singh, G. Poornalatha, and S. Singh, "BREE-HD: A transformer-based model to identify threats on Twitter," *IEEE Access*, vol. 11, pp. 67180–67190, 2023.

[7] Y. Gao, Y. Kim, B. G. Doan, Z. Zhang, G. Zhang, S. Nepal, D. C. Ranasinghe, and H. Kim, "Design and evaluation of a multi-domain trojan detection method on deep neural networks," *IEEE Trans. Depend. Secure Comput.*, vol. 19, no. 4, pp. 2349–2364, Jul. 2022.

[8] I. Aliyu, S. Van Engelenburg, M. B. Mu'Azu, J. Kim, and C. G. Lim, "Statistical detection of adversarial examples in blockchain-based federated forest in-vehicle network intrusion detection systems," *IEEE Access*, vol. 10, pp. 109366–109384, 2022.

[9] K. Gu, X. Dong, X. Li, and W. Jia, "Cluster-based malicious node detection for false downstream data in fog computing-based VANETs," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 3, pp. 1245–1263, May 2022.

[10] T. T. Huong, T. P. Bac, K. N. Ha, N. V. Hoang, N. X. Hoang, N. T. Hung, and K. P. Tran, "Federated learning-based explainable anomaly detection for industrial control systems," *IEEE Access*, vol. 10, pp. 53854–53872, 2022.

[11] G.-Y. Shin, D.-W. Kim, and M.-M. Han, "Data discretization and decision boundary data point analysis for unknown attack detection," *IEEE Access*, vol. 10, pp. 114008–114015, 2022.

[12] E. Paolini, L. Valcarenghi, L. Maggiani, and N. Andriolli, "Real-time clustering based on deep embeddings for threat detection in 6G networks," *IEEE Access*, vol. 11, pp. 115827–115835, 2023.

[13] F. Rustam, A. Raza, M. Qasim, S. K. Posa, and A. D. Jurcut, "A novel approach for real-time server-based attack detection using meta-learning," *IEEE Access*, vol. 12, pp. 39614–39627, 2024.

[14] U. Sabeel, S. S. Heydari, K. El-Khatib, and K. Elgazzar, "Incremental adversarial learning for polymorphic attack detection," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 2, pp. 869–887, 2024.

[15] C. A. Fadhilla, M. D. Alfikri, and R. Kaliski, "Lightweight meta-learning BotNet attack detection," *IEEE Internet Things J.*, vol. 10, no. 10, pp. 8455–8466, May 2023.

[16] H. Whitworth, S. Al-Rubaye, A. Tsourdos, and J. Jiggins, "5G aviation networks using novel AI approach for DDoS detection," *IEEE Access*, vol. 11, pp. 77518–77542, 2023.

[17] Z. A. E. Houda, D. Naboulsi, and G. Kaddoum, "A privacy-preserving collaborative jamming attacks detection framework using federated learning," *IEEE Internet Things J.*, vol. 11, no. 7, pp. 12153–12164, Apr. 2024.

[18] A. B. Bhutto, X. S. Vu, E. Elmroth, W. P. Tay, and M. Bhuyan, "Reinforced transformer learning for VSI-DDoS detection in edge clouds," *IEEE Access*, vol. 10, pp. 94677–94690, 2022.

[19] M. Wazid, J. Singh, A. K. Das, and J. J. P. C. Rodrigues, "An ensemble-based machine learning-envisioned intrusion detection in industry 5.0-driven healthcare applications," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 1903–1912, Feb. 2024.

[20] Y.-W. Chang, H.-Y. Shih, and T.-N. Lin, "AI-URG: Account identity-based uncertain graph framework for fraud detection," *IEEE Trans. Computat. Social Syst.*, vol. 11, no. 3, pp. 3706–3728, Jun. 2024.

[21] N. Moustafa, K. R. Choo, and A. M. Abu-Mahfouz, "Guest editorial: AI-enabled threat intelligence and hunting microservices for distributed industrial IoT system," *IEEE Trans. Ind. Informat.*, vol. 18, no. 3, pp. 1892–1895, Mar. 2022.

[22] O. Toker, "Asymptotic performance limitations in cyberattack detection," *IEEE Open J. Circuits Syst.*, vol. 4, pp. 336–346, 2023.

[23] G. Bendiab, A. Hameurlaine, G. Germanos, N. Kolokotronis, and S. Shiaeles, "Autonomous vehicles security: Challenges and solutions using blockchain and artificial intelligence," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 4, pp. 3614–3637, Apr. 2023.

[24] G. Ahmadi-Assalemi, H. Al-Khateeb, G. Epiphaniou, and A. Aggoun, "Super learner ensemble for anomaly detection and cyber-risk quantification in industrial control systems," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 13279–13297, Aug. 2022.

[25] F. S. Alsubaei, A. A. Almazroi, and N. Ayub, "Enhancing phishing detection: A novel hybrid deep learning framework for cybercrime forensics," *IEEE Access*, vol. 12, pp. 8373–8389, 2024.

[26] P. Lachkov, L. Tawalbeh, and S. Bhatt, "Vulnerability assessment for applications security through penetration simulation and testing," *J. Web Eng.*, vol. 21, pp. 2187–2208, Dec. 2022.

[27] M. Torres, R. Álvarez, and M. Cazorla, "A malware detection approach based on feature engineering and behavior analysis," *IEEE Access*, vol. 11, pp. 105355–105367, 2023.
[28] R. Dubin, "Disarming attacks inside neural network models," *IEEE Access*, vol. 11, pp. 124295–124303, 2023.
[29] E. Paolini, L. Valcarenghi, L. Maggiani, and N. Andriolli, "Real-time network packet classification exploiting computer vision architectures," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 1155–1166, 2024.
[30] M. A. Taher, M. Behnamfar, A. I. Sarwat, and M. Tariq, "False data injection attack detection and mitigation using nonlinear autoregressive exogenous input-based observers in distributed control for DC microgrid," *IEEE Open J. Ind. Electron. Soc.*, vol. 5, pp. 441–457, 2024.
[31] R. Allafi and I. R. Alzahrani, "Enhancing cybersecurity in the Internet of Things environment using artificial orca algorithm and ensemble learning model," *IEEE Access*, vol. 12, pp. 63282–63291, 2024.
[32] C.-C. Chang, H. H. Nguyen, J. Yamagishi, and I. Echizen, "Cyber vaccine for deepfake immunity," *IEEE Access*, vol. 11, pp. 105027–105039, 2023.
[33] D. Namakshenas, A. Yazdinejad, A. Dehghantanha, R. M. Parizi, and G. Srivastava, "IP2FL: Interpretation-based privacy-preserving federated learning for industrial cyber-physical systems," *IEEE Trans. Ind. Cyber-Phys. Syst.*, vol. 2, pp. 321–330, 2024.
[34] Z. Rahman, X. Yi, and I. Khalil, "Blockchain-based AI-enabled Industry 4.0 CPS protection against advanced persistent threat," *IEEE Internet Things J.*, vol. 10, no. 8, pp. 6769–6778, Apr. 2023.
[35] S. Neupane, J. Ables, W. Anderson, S. Mittal, S. Rahimi, I. Banicescu, and M. Seale, "Explainable intrusion detection systems (X-IDS): A survey of current methods, challenges, and opportunities," *IEEE Access*, vol. 10, pp. 112392–112415, 2022.
[36] H. Suryotrisongko, Y. Musashi, A. Tsuneda, and K. Sugitani, "Robust botnet DGA detection: Blending XAI and OSINT for cyber threat intelligence sharing," *IEEE Access*, vol. 10, pp. 34613–34624, 2022.
[37] K. A. Abuhasel, "A linear probabilistic resilience model for securing critical infrastructure in Industry 5.0," *IEEE Access*, vol. 11, pp. 80863–80873, 2023.
[38] M. Qasim, M. Waleed, T.-W. Um, P. Pahlevani, J. M. Pedersen, and A. Masood, "Diving deep with BotLab-DS1: A novel ground truth-empowered botnet dataset," *IEEE Access*, vol. 12, pp. 28898–28910, 2024.
[39] P. Verma, N. Bharot, J. G. Breslin, D. O'Shea, A. Vidyarthi, and D. Gupta, "Zero-day guardian: A dual model enabled federated learning framework for handling Zero-day attacks in 5G enabled IIoT," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 3856–3866, Feb. 2024.
[40] M. Ali, G. Kaddoum, W.-T. Li, C. Yuen, M. Tariq, and H. V. Poor, "A smart digital twin enabled security framework for vehicle-to-grid cyber-physical systems," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 5258–5271, 2023.
[41] A. Guesmi, M. A. Hanif, B. Ouni, and M. Shafique, "SAAM: Stealthy adversarial attack on monocular depth estimation," *IEEE Access*, vol. 12, pp. 13571–13585, 2024.
[42] S. Gupta, C. Maple, and R. Passerone, "An investigation of cyber-attacks and security mechanisms for connected and autonomous vehicles," *IEEE Access*, vol. 11, pp. 90641–90669, 2023.
[43] D. He, X. Lv, X. Xu, S. Chan, and K.-K.-R. Choo, "Double-layer detection of internal threat in enterprise systems based on deep learning," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 4741–4751, 2024.
[44] T. A. Al-Shehari, D. Rosaci, M. Al-Razgan, T. Alfakih, M. Kadrie, H. Afzal, and R. Nawaz, "Enhancing insider threat detection in imbalanced cybersecurity settings using the density-based local outlier factor algorithm," *IEEE Access*, vol. 12, pp. 34820–34834, 2024.
[45] S.-H. Park, S.-W. Yun, S.-E. Jeon, N.-E. Park, H.-Y. Shim, Y.-R. Lee, S.-J. Lee, T.-R. Park, N.-Y. Shin, M.-J. Kang, and I.-G. Lee, "Performance evaluation of open-source endpoint detection and response combining Google rapid response and osquery for threat detection," *IEEE Access*, vol. 10, pp. 20259–20269, 2022.
[46] I. Chen, L. Huang, J. Qiao, D. E. Tamir, and N. Rishe, "Combining perception considerations with artificial intelligence in maritime threat detection systems," in *Proc. 17th Annu. Syst. Syst. Eng. Conf. (SOSE)*, Jun. 2022, pp. 417–422.
[47] K. T. Nitesh, A. K. Thirumala, U. F. Mohammed, and M. R. Ahmed, "Network security threat detection: Leveraging machine learning algorithms for effective prediction," in *Proc. 12th Int. Conf. Adv. Comput. (ICoAC)*, Aug. 2023, pp. 1–5.
[48] T. B. Ghuge and S. S. Biradar, "Web data mining for cyber security threat detection," in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, Apr. 2024, pp. 1420–1426.

KAVITHA DHANUSHKODI received the Master of Engineering and Ph.D. degrees in computer science and engineering from Anna University, Chennai. She is currently an Associate Professor with the School of Computer Science and Engineering(SCOPE), Vellore Institute of Technology, Chennai Campus, Chennai, Tamil Nadu, India. She has an overall teaching experience of 16 years in various academic institutions. She has published more than 42 research articles to her credit in reputed journals. Her research interests include software security, the Internet of Things, and cyber security.

S. THEJAS is currently pursuing the M.Tech. degree in computer science and engineering from Vellore Institute of Technology, Chennai, with a focus on integrating artificial intelligence in cybersecurity to mitigate evolving cyber threats. He is a Cybersecurity Enthusiast and a Researcher with expertise in AI-enabled threat detection and advanced security solutions. He has completed an internship focused on ethical hacking and smartphone sensor configuration, gaining hands-on experience with advanced tools in Kali Linux and AI-driven security frameworks. His research interests include AI-driven intrusion detection, explainable AI, and the application of machine learning in enhancing cybersecurity resilience.