Physical Education, Health and Social Sciences

https://journal-of-social-education.org

E-ISSN: 2958-5996

P-ISSN: 2958-5988

# AI-Driven Cybersecurity: Threat Detection, Prevention, and Autonomous Defense

**Syeda Hina Shah*[1], Kamal Khan[2], Norin Salim[3], Engr. Muhammad Anwar Reki[4], Summayya Shabbir Baloch[5], Ghulam Yasin[6]**

[1] Harbin Institute of Technology, Shenzhen, China.
Corresponding Author: engrsyeda8@gmail.com

[2] Department of Computer Science, University of Makran. Kamalkhan@uomp.edu.pk

[3] Department of Computer Networks and Security, National University of Computer and Emerging Sciences. norinsalim@gmail.com

[4] Department of Computer Science, University of Makran, Panjgur.
muhammadanwar@uomp.edu.pk

[5] Department of Computer Science, University of Makran, Panjgur.
summayyabaloch634@gmail.com

[6] Department of Computer science, University of Makran, Panjgur.
yasinarman38@gmail.com

**DOI: https://doi.org/10.63163/jpehss.v3i4.955**

**Abstract**

The rapid expansion of digital ecosystems has intensified cyber threats, exposing the limitations of traditional, signature-based security systems. Artificial Intelligence (AI) has emerged as a transformative enabler of advanced cyber defense, offering adaptive, scalable, and autonomous security capabilities. This review provides a comprehensive synthesis of AI-driven cybersecurity mechanisms across three core defense layers**,** threat detection, proactive prevention, and autonomous response. Deep learning architectures including CNNs, LSTMs, GRUs, hybrid CNN-BiLSTM models, and self-normalizing networks have significantly improved intrusion detection accuracy and reduced false positives. Predictive Vulnerability Exploitation (PVE) models, such as EPSS, enhance vulnerability prioritization by quantifying real-world exploit likelihood. Autonomous defense frameworks, powered by Deep Reinforcement Learning (DRL) and agent-based Large Language Models (LLMs), enable zero-day attack detection, dynamic playbook generation, and zero-shot incident response. However, challenges such as adversarial machine learning, model poisoning, bias, privacy concerns, supply-chain insecurity, and dual-use risks remain substantial barriers to trustworthy deployment. Future directions emphasize federated learning, privacy-preserving intelligence sharing, post-quantum security integration, and neuromorphic hardware for ultra-low-latency edge defense. Overall, AI has shifted cybersecurity from reactive monitoring to predictive and autonomous protection, marking a foundational transformation in digital defense ecosystems.

**Keywords:** Artificial Intelligence; Cybersecurity; Threat Detection; Intrusion Detection Systems; Deep Learning; Reinforcement Learning; Predictive Vulnerability Exploitation; Autonomous Defense; SOAR; Generative AI; Adversarial Machine Learning; Federated Learning; Post-Quantum Security.

## I. Introduction

The rapid pace of the digital revolution has introduced an era of unprecedented cyber risks, necessitating a dramatic move beyond conventional security methodologies Cyberattacks have become increasingly sophisticated, targeting critical infrastructure, sensitive systems, and individuals to compromise data and disrupt essential operations (Adelusola et al, 2024). Traditional cybersecurity measures, while foundational, are often reactive, labor-intensive, and inherently struggle to cope with the complexity, high dimensionality, and dynamic nature of modern threats, particularly those involving zero-day vulnerabilities (Altunay et al., 2023). In this context, Artificial Intelligence (AI) has emerged as a crucial and "game-changing technology" capable of fundamentally enhancing the security framework of network environments (Carahsoft et al, 2025). The integration of AI has demonstrably increased the ability to identify and counteract advanced threats, including adversarial assaults and network breaches (Darmadi et al, 2025). This review provides a systematic analysis of current research, emphasizing AI's application across the entire cybersecurity lifecycle (FIRST et al, 2025). The domains of focus correspond to the progression of cyber defense: Threat Detection (identifying intrusions in real-time), Proactive Prevention (predicting and prioritizing vulnerabilities), and Autonomous Defense (adaptive, self-adjusting incident response) (ISC2 et al, 2024).

The successful deployment of AI hinges on its trustworthiness, which extends beyond high detection rates to include assurances of model explainability and resilience. As systems shift from merely augmenting human analysts to autonomous decision-makers, the reliability of AI outputs becomes a prerequisite for operational success. (Polemi et al 2024). This transition signals a fundamental paradigm change: the primary value of AI is not solely faster identification, but the capacity for delivering responsive, scalable, and adaptive countermeasures at machine speed (KPMG et al,. 2025). Machine Learning (ML) Used broadly for tasks such as behavioral analysis, static file analysis, and anomaly detection to inform risk scoring and guide threat investigations. Deep Learning (DL) Utilizes multi-layered neural networks (e.g., CNNs, RNNs, LSTMs) to process complex, high-dimensional security data, particularly in high-fidelity intrusion detection systems (IDS) (Li, M,. 2017). Deep Reinforcement Learning (DRL) By incorporating deep neural networks into traditional Reinforcement Learning (RL), DRL is highly capable of solving dynamic defense problems and enabling autonomous decision-making and optimal response strategies (Munikoti,. et al,. 2023). Generative AI (GenAI) and Large Language Models (LLMs) These are revolutionizing security automation by generating synthetic attack data to enhance threat simulation and powering autonomous agents for dynamic playbook creation and execution (Luo et al., 2025).

**Figure 1.1 Key Applications of Artificial Intelligence in Modern Cybersecurity**

## 2. AI-Driven Threat Detection Systems

### Deep Learning Architectures for Network Intrusion Detection (NIDS)

Network Intrusion Detection Systems (NIDS) are critical for monitoring network traffic. These systems traditionally rely on signature-based methods, but Deep Learning (DL) models significantly enhance anomaly detection capabilities, which are essential for identifying previously unseen attack patterns (Zhang et al., 2024). DL approaches are utilized to analyze vast volumes of data, searching for patterns and indicators of compromise. The effectiveness of these models in improving efficiency and rapid response is continuously demonstrated, with some deep learning architectures achieving classification accuracies of 96 percent, underscoring their potential for enhancing adaptive IDS solutions (Zhou & Yan, 2025). The deep learning models applied in NIDS research are diverse, including Deep Belief Networks (DBNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long-Short-Term-Memory (LSTM), Gated Recurrent Units (GRUs), and autoencoders (Li et al., 2024). These architectures are tailored to different tasks, such as classifying malicious features or identifying deviations from typical user behavior (Memon, N et al, 2024).

### Performance Evaluation on Benchmark Datasets

Recurrent architectures are particularly valuable for extracting temporal features from sequential network traffic data. In comparative studies using the widely recognized NSL-KDD dataset, the GRU approach has demonstrated exceptional capabilities in binary classification tasks, often outperforming other recurrent models (Li et al., 2024). Similarly, the combination of XGBoost with Simple-RNN or LSTM proves efficient for multi-class classification on benchmark datasets like NSL-KDD and UNSW-NB15 (NIST et al,. 2025). High-fidelity detection is often achieved through hybrid models that combine multiple architectural strengths. A Hybrid CNN-Bi-LSTM architecture, which extracts both spatial and temporal features, has shown outstanding results, achieving reported accuracies of 99.87% on the UNSW-NB15 binary dataset (Sahu, A et al,. 2024). This efficacy relies on sophisticated feature engineering and the integration of advanced concepts, such as Transformer modules, to enhance expressiveness and consider global relationships in network data (Salehi, S et al, 2023). Despite the high reported accuracy in academic benchmarks, a major operational consideration for anomaly-based NIDS is the propensity for high False Alarm Rates (FAR) (Seemplicity et al, 2025). This trade-off means that models optimized solely for high recall may generate excessive alerts, leading to alert fatigue in Security Operations Centers (SOCs). Consequently, practical deployment favors novel architectures designed specifically for robustness and low False-Positive Rates (FPR), such as the Self-Normalizing Neural Network (SNN)-based DeepDCA model, which exhibits superior performance in detecting IoT attacks compared to conventional ML classifiers (Zhang et al., 2024).

### Malware and Endpoint Protection

Machine learning enables organizations to automate threat detection and response by augmenting traditional signature-based methods with a generalized approach that rapidly learns the difference between benign and malicious samples (Silver, D et al,. 2017). ML processes are divided into several key analysis types for endpoint security: static file analysis, behavioral analysis, and hybrid analysis, which combines both for advanced threat detection (Wang, Y et al,. 2025).

Deep learning models, however, suffer from inherent challenges such as limited adaptability and susceptibility to adversarial manipulation. This is exacerbated by the difficulty of obtaining balanced, real-world datasets of new attack vectors. Generative Adversarial Networks (GANs) offer a powerful solution to this "data problem" by generating synthetic data that mimics real-world attack patterns, enabling robust training simulations (Yan et al., 2025). This capability

| Model Architecture | Dataset Benchmark | Classification Type | Reported Accuracy (%) | Key Architectural Advantage | Reference |
|---|---|---|---|---|---|
| **Hybrid CNN-Bi-LSTM** | UNSW-NB15 | Binary | 99.87% | Superior Spatiotemporal Feature Extraction | (Altunay & Albayrak, 2023) |
| **GRU** | NSL-KDD | Binary | Outperforms others | Efficacy in time series analysis and anomaly detection | (Li et al., 2024) |
| **DeepDCA (SNN)** | IoT Attacks | Binary | High Accuracy, Low FPR | Novel Self-Normalizing Technique for Robustness | (Zhang et al., 2024) |
| **XGBoost-Simple-RNN** | UNSW-NB15 | Binary | 87.07% | High efficiency and competitive performance | (Zhang et al., 2024) |
| **DNN** | NSL-KDD | Multi-class | Outstanding Performance | Most effective in handling complex network patterns | (Li et al., 2024) |

transforms the defense landscape by simulating sophisticated attack scenarios, improving adversarial robustness, and mitigating dataset imbalance in domains like malware detection. (Munikoti et al., 2023). However, this capability is dual-use; attackers also exploit GANs to create potent adversarial malware examples. Research on models like Mal-D2GAN demonstrates that advanced GAN architectures can successfully reduce the detection accuracy of existing malware detectors, emphasizing the constant arms race in model robustness (Zhou, M et al,. 2025).

### 3. Proactive AI for Threat Prevention and Mitigation

#### Predictive Vulnerability Exploitation (PVE) Models

Proactive prevention shifts the defense posture from reacting to incidents to predicting and preemptively neutralizing threats (Seemplicity, 2025). AI analyzes historical data and previous security breaches to predict cyberattacks and stay ahead of emerging threats, fundamentally improving vulnerability management practices (Adelusola et al., 2024). A key application is dynamic risk scoring. Machine learning evaluates threats, vulnerabilities, and misconfigurations, assigning numerical risk scores that dynamically adjust based on asset sensitivity and exploitability, which then drive intelligent remediation efforts. This paradigm moves the focus of vulnerability management from theoretical maximum severity to *imminent threat likelihood* (Altunay et al., 2023). The Exploit Prediction Scoring System (EPSS) is a widely adopted machine learning model that operationalizes this shift. EPSS estimates the likelihood (probability between 0 and 1, or 0% and 100%) that a known software vulnerability will be exploited in the wild within the next 30 days (FIRST, 2025). Unlike traditional metrics that measure severity, EPSS uses

vulnerability information and real-world exploitation activity (CVE data) to measure *threat*, providing a crucial tool for network defenders to prioritize remediation efforts effectively. The model, currently in Version 4, is utilized to produce daily estimates for all published CVEs (ISC2 et al,. 2024). Practical vulnerability management is achieved by integrating these AI-derived risk scores with catalogs of actively exploited issues, such as the CISA Known Exploited Vulnerabilities (KEV) list (CISA et al,. 2024). PVE models often utilize hybrid neural network architectures, such as combinations of RNNs and LSTMs, to process sequential vulnerability data (Darmadi et al,. 2025). These deep learning frameworks identify low-dimensional distributed representations (embeddings) derived from threat intelligence, including data sources like the Dark Web, to predict vulnerable application components (Memon, N et al,. 2024).

**AI in Policy Enforcement and User Behavior Analytics (UBA)**

Machine learning serves as a critical force-multiplier in security operations, automating repetitive manual tasks and redirecting resources toward complex, strategic projects (Li, M,. 2017). AI strengthens defenses by continuously monitoring systems to isolate compromised devices and block malicious traffic (Sahu, A et al,. 2024). For policy enforcement, AI tools analyze user authentication data such as fingerprints, typing styles, or voice patterns and behavioral patterns to authenticate users and identify unusual activity that may signal a compromise or insider threat (Salehi, S et al,. 2023). Generative AI capabilities further enable preemptive defense by analyzing vast datasets of past attacks to predict potential future attack scenarios, enhancing threat detection models through the generation of synthetic data that mimics real-world threats (Seemplicity et al,. 2025).

**The ML Supply Chain as a New Threat Vector**

While AI enhances security, the rapid adoption of Machine Learning introduces new security concerns related to the development supply chain. ML engineers frequently download vast, often unmonitored, dependencies and packages for development environments (Silver, D et al., 2017). This focus on experimentation velocity over security creates a "chaotic and loosely controlled ecosystem" where vulnerable dependencies, often containing Arbitrary Code Execution (ACE) flaws, are overlooked (Wang, Y et al,. 2025). This exposes the environment to resurging attack vectors, such as drive-by attacks, which are triggered by the mere inclusion of vulnerable dependency. This situation necessitates proactive prevention strategies extend to rigorously securing the entire ML development supply chain, an often-neglected threat surface (Zhou, M et al,. 2025).

**4. Autonomous Defense and Adaptive Response**

**Deep Reinforcement Learning (DRL) for Zero-Day Defense and Adaptive Mitigation**

True autonomous defense relies on mechanisms that are responsive, adaptive, and scalable against the most dynamic and complex threats. Deep Reinforcement Learning (DRL) fulfills this role by providing the computational framework to solve high-dimensional defense problems (Salehi & Mohajer, 2023). The DRL process involves an agent, often using algorithms like Deep Q-Networks (DQN) or Proximal Policy Optimization (PPO), which interacts with the network environment, taking actions and receiving a signal reward based on the correctness of its predictions (Li, 2017; Sahu, 2024). The agent learns an optimal policy to maximize its cumulative reward over time. This approach abstracts cyber conflict as an adversarial game, enabling DRL to move beyond simple detection toward strategic, continuous optimization of countermeasures against evolving threats (Silver, et al., 2017). DRL-based Network Intrusion Detection Systems (NIDS), often utilizing stacked LSTM architectures, are specifically designed for Zero-Day attack detection (Darmadi &

Saputra, 2025). This involves training the agent on known attacks but withholding specific attack categories from the training dataset to test the agent's ability to identify entirely unseen threats. Data balancing techniques, such as K-means SMOTE and ADASYN, are crucial to maintaining performance amidst the class imbalance inherent in zero-day attack datasets. Furthermore, DRL, particularly Deep Q-Networks (DQN), is leveraged for real-time zero-day vulnerability detection by discovering underlying patterns and features in telemetry data without prior knowledge of the vulnerabilities, demonstrating scalability to complex network state spaces (Sahu, 2024).

| DRL Algorithm | Cybersecurity Task | Mechanism/Architecture | Proactive Capability | Reference |
|---|---|---|---|---|
| DRL (Stacked LSTM) | Zero-Day Attack Detection (NIDS) | Oversampling Techniques (SMOTE/ADASYN) | Identification of previously unseen attack patterns | (Darmadi & Saputra, 2025) |
| Deep Q-Networks (DQN) | Zero-Day Vulnerability Detection | Learning without prior vulnerability knowledge, Reward Maximization | Real-time adaptation and zero-day pattern discovery | (Sahu, 2024) |
| DRL (PPO) | Adversarial Simulation/Defense | Proximal Policy Optimization, Adversarial Environment | Learning optimal countermeasure sequences | (Li, 2017) |
| LLM Agents (IRCopilot) | Autonomous Incident Response (SOAR) | Planner-Analyst-Generator Framework | Zero-shot task execution and dynamic playbook generation | (Wang, Li, Zhang, & Chen, 2025) |

**Agentic Security Orchestration, Automation, and Response (SOAR)**

AI-powered SOAR platforms are revolutionizing incident response by transitioning from reactive, predefined automation to proactive, self-adjusting threat mitigation (Memon, N et al,. 2024). Novel frameworks utilize autonomous Large Language Model (LLM) agents, capable of dynamically generating, adapting, and executing contextual playbooks. This capacity for "zero-shot task execution" allows the system to construct novel response sequences for unique incidents, overcoming the limitations of static automation (Luo et al., 2025). Technology leverages AI to autonomously evaluate threats, make informed decisions, and initiate responses in real time without human intervention (NIST et al,. 2025).

Advanced LLMs, such as GPT 4 or Gemini, power these multi-agent systems (MAS), enabling them to perform complex tasks autonomously (Salehi, S et al,. 2023). Each AI agent is typically specialized for a particular stage of investigation, triage, or response (Seemplicity et al,. 2025). The IRCopilot framework, for instance, demonstrates the efficacy of a multi-step reasoning agent (comprising a Planner, Analyst, and Generator) in overcoming the challenges of generating the correct response sequence (Wang, Li, Zhang, & Chen, 2025). Experimental results show IRCopilot achieves significantly higher sub-task completion rates—up to 150% improvement over baseline LLMs in various response tasks (Wang et al., 2025). The substantial gain in performance results from the ability of LLMs to provide contextual reasoning, allowing agents to interpret complex

alerts and dynamically orchestrate external tool integration based on the unique situation of the incident (Luo et al., 2025).

**Self-Healing and Adaptive Network Architectures**

Self-healing networks represent the pinnacle of autonomous defense, employing AI and machine learning for rapid recovery and autonomous network management following disruption (Silver, D et al,. 2017). This paradigm shifts security from passive defense to proactive protection. Reinforcement learning is applied to continuously evaluate and refine defense mechanisms by simulating cyberattacks and optimizing response strategies dynamically, allowing for the mitigation of zero-day threats before human intervention is required (Wang, Y et al,. 2025). This ensures critical infrastructure and core services remain operational despite network issues (Yan, Z et al,. 2025).

Figure 4.1 Core Components of AI-Driven Threat Hunting in Proactive Cyber Defense


## 5. Critical Challenges and Dual-Use Risks in AI Cybersecurity

**Adversarial Machine Learning (AML)**

Adversarial Machine Learning (AML) constitutes a primary strategic vulnerability for AI driven security systems. AML involves malicious techniques that deliberately manipulate machine learning models by feeding them deceptive data to induce incorrect or unintended behavior (Adelusola et al., 2024). Evasion Attacks These attacks occur during the inference stage, where attackers create "adversarial examples" inputs with subtle, imperceptible alterations to fool a trained model and bypass defenses (Altunay et al., 2023). This type of attack is particularly insidious because it can lead to the silent, long-term degradation of a model's performance and compromise decision-making processes without immediate human detection (ISC2 et al,. 2024).

Poisoning Attacks These attacks target the training data, deliberately introducing bias or corrupting the dataset to skew the model’s outputs and degrade security posture. Common methods include data injection, mislabeling, and backdoor poisoning (Darmadi et al,. 2025).

Model Extraction This is listed under the NIST taxonomy of attacks on Predictive AI (PredAI) systems, involving the replication of a proprietary model's underlying logic through query access, which can facilitate further evasion attacks (National Institute of Standards and Technology, 2025).

Defense against AML utilizes reciprocal techniques, such as red teaming and ethical hacking, to exploit models defensively and develop stronger, more robust algorithms. Architectural

improvements, such as the Mal-D2GAN, aim to enhance resilience against maliciously crafted inputs (Wang et al., 2025).

**Ethical, Regulatory, and Operational Bottlenecks**

The widespread adoption of AI in cyber defense raises significant ethical and governance challenges that must be addressed concurrently with technological advancement (Adelusola & Adelusola, 2024). A core ethical conflict is the tension between security and privacy. The analytical capability of AI systems relies on processing vast amounts of personal and network data, creating concerns regarding surveillance, data collection, and the potential for misuse that infringes upon civil liberties. Leaders must deploy AI responsibly to enhance security without compromising individual privacy rights (KPMG, 2025).

Furthermore, data poisoning or inherent systemic issues can amplify algorithmic bias and fairness problems. An AI-based system trained with biased data might disproportionately flag legitimate software used by specific demographics as malicious, leading to unfair profiling, discrimination, and unjust consequences (ISC2, 2024). The autonomy of systems introduces profound questions of accountability and transparency. Autonomous response systems are highly effective (Sahu, A et al,. 2024), but their complexity complicates the process of understanding decision-making, particularly when errors occur (Memon et al., 2024). The regulatory challenge lies in providing governance, such as the NIST AI Risk Management Framework, to ensure that human oversight and managerial accountability remain crucial, mitigating legal liabilities associated with unpredictable autonomous actions (Seemplicity et al,. 2025).

Finally, the dual-use dilemma dictates that AI technology is equally available for offensive use, enabling attackers to scale and sophisticate cyberattacks. Additionally, the automation of routine tasks presents the ethical dilemma of job displacement within the cybersecurity industry, requiring proactive strategies for retraining and reskilling the workforce (ISC2, 2024).

## 5. Future Directions and Emerging Paradigms

**Collaborative Intelligence and Privacy-Preserving Techniques**

As cyber threats become increasingly transnational, organizations face escalating challenges in sharing sensitive threat intelligence due to data confidentiality requirements and geopolitical boundaries. Traditional centralized models struggle to overcome these barriers (Zhang, Q et al,. 2024). Federated Learning (FL) offers a paradigm shift for collaborative cybersecurity. FL is a decentralized machine learning approach that trains a shared model across multiple entities holding local data samples, transmitting only model updates, not raw data (Salehi, S et al,. 2023). This approach preserves data sovereignty and minimizes the risk of data breaches during transfer, making it ideal for multinational corporations and governmental bodies operating under diverse data privacy laws. By enabling collaborative threat intelligence across these organizational and geopolitical boundaries, FL represents a mechanism for strategic, global defense (Seemplicity et al,. 2025). However, FL implementations must integrate advanced cryptographic safeguards, such as differential privacy and homomorphic encryption, to mitigate specific risks like model poisoning and inference attacks orchestrated by adversarial participants (Wang, Y et al,. 2025).

**Post-Quantum Security Integration**

The threat posed by quantum computing, specifically its potential to break current public-key encryption, necessitates urgent action toward cryptographic agility. The "Harvest Now, Decrypt Later" strategy, where encrypted data is accumulated for future decryption by quantum technologies, heightens the urgency for adopting Post-Quantum Cryptography (PQC) (Li, M,. 2017). PQC will quickly become an immediate compliance and business continuity issue (Altunay

et al., 2023). AI plays a key role in this transition. First, AI-powered preemptive security will become standard, enabling Autonomous Security Operations Centres (SOCs) to anticipate and neutralize threats in milliseconds. Second, AI-driven detection and adaptive defense are critical complements to PQC migration (Carahsoft et al,. 2025). AI models establish behavioral baselines and provide continuous real-time monitoring and trust validation, ensuring strong, future-proof encryption standards and protecting sensitive data integrity against quantum-era decryption risks (Luo, J et al,. 2025).

**Hardware and Architectural Innovations**

The ambition to achieve autonomous defense systems capable of neutralizing threats in milliseconds is fundamentally constrained by the power consumption and latency of traditional GPU/CPU architectures (Silver, D et al,. 2017). Emerging Neuromorphic Computing architectures are poised to address this hardware bottleneck (CISA et al,. 2024). Neuromorphic hardware utilizes event-driven processors optimized for ultra-low power consumption, executing the core computational model of the Spiking Neural Network (SNN). SNNs encode information through the timing and frequency of spikes, drastically reducing computational overhead. This architectural shift enables real time AI inference and learning directly at the network edge, providing the necessary low latency and energy efficiency to move threat detection capabilities closer to the source of the traffic, thus maximizing the speed and ubiquity of autonomous defense systems (Santoso et al,. 2024).

## 6. Conclusion

AI has catalyzed a fundamental transformation in cybersecurity, shifting the paradigm from static, reactive measures to dynamic, predictive, and autonomous defense frameworks. Critical advancements include the near-perfect accuracy achieved by hybrid deep learning models in intrusion detection (Altunay & Albayrak, 2023), the maturation of Predictive Vulnerability Exploitation (PVE) models like EPSS for intelligence-driven prioritization (FIRST, 2025), and the emergence of Deep Reinforcement Learning (DRL) agents (Darmadi & Saputra, 2025) and Agentic LLM-based SOAR for autonomous, self-adjusting incident response (Luo et al., 2025; Wang et al., 2025). These innovations are essential for maintaining a technological edge and scaling human capabilities against increasingly sophisticated and rapid cyber threats (Carahsoft, 2025). To fully realize the potential of AI-driven security, future efforts must prioritize strategic and policy-driven advancements. This includes developing robust defenses against Adversarial Machine Learning (AML) to prevent the silent degradation of model, establishing transparent and accountable governance frameworks (like the NIST AI Risk Management Framework) for autonomous systems (Carahsoft, 2025; National Institute of Standards and Technology, 2025), and accelerating the adoption of collaborative, privacy-preserving technologies like Federated Learning (FL) and future-proof cryptographic standards (PQC) integrated with specialized hardware architectures (Memon, N et al,. 2024).

**Acknowledgment**

We would like to express my deepest gratitude to my parents for their unconditional love, constant encouragement, and unwavering support throughout my academic journey. Their guidance and prayers have always been a source of strength and motivation for me. We are also sincerely thankful to my teachers, whose dedication, knowledge, and mentorship have shaped my learning and inspired me to strive for excellence. Their continuous support and valuable guidance played an essential role in the completion of this work.

## References


Adelusola, M., & Adelusola, D. (2024). *AI-Driven Cybersecurity: A Systematic Review of Current Research and Future Directions*. ResearchGate.

Altunay, H. C., & Albayrak, Z. (2023). A hybrid CNN+ LSTM-based intrusion detection system for industrial IoT networks. *Engineering Science and Technology, an International Journal*, *38*, 101322.

Carahsoft. (2025). *Making the Best Use of AI*.

CISA. (2024). *CISA Artificial Intelligence Use Cases*.

Darmadi, A., & Saputra, A. A. (2025). Adaptive Defense: Zero-Day Attack Detection in NIDS With Deep Reinforcement Learning. *IEEE Access*, *13*, 116345–116361.

FIRST. (2025). *Exploit Prediction Scoring System (EPSS)* (Version 4).

ISC2. (2024). *The Ethical Dilemmas of AI in Cybersecurity*.

KPMG. (2025). *The ethical use of AI in cybersecurity*.

Li, M. (2017). *Reinforcement learning for cyber defense*. AAAI.

Li, N., Liu, T., & Liu, P. (2024). A Comparative Study of Using Deep Learning Algorithms in Network Intrusion Detection. *IEEE Access*, *12*, 58851–58870.

Luo, J., Chen, T., Zhang, Y., & Li, F. (2025). Agentic LLM-Based SOAR Architecture: Dynamic Playbook Generation and Tool Orchestration for Proactive Threat Mitigation. *Informatics*, *16*(5), 365.

Memon, N., Chen, T., & Wang, Y. (2024). AI Enabled Threat Detection: Leveraging Artificial Intelligence for Advanced Security and Cyber Threat Mitigation. *IEEE Access*, *12*, 173127–173136.

National Institute of Standards and Technology. (2025). *NIST AI 100-2e2025: Adversarial Machine Learning (AML) on Predictive AI Systems*.

Sahu, A. (2024). A novel approach for zero-day vulnerability detection using Deep Q-Networks. *PLOS ONE*, *19*(5), e0324595.

Salehi, S., & Mohajer, M. (2023). Deep Reinforcement Learning for Cyber Security. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(8), 3779–3795.

Seemplicity. (2025). *F_DR_Seemplicity_State-of-Report_March2025*.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A.,... & Chen, Y. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–359.

Wang, X., Li, X., Zhang, Y., & Chen, J. (2025). IRCopilot: Zero-Shot Autonomous Incident Response with Large Language Model Agents. *arXiv:2505.20945*.

Wang, Y., Li, Z., & Chen, H. (2025). Mal-D2GAN: Adversarial Malware Generation Using Double-Detector Generative Adversarial Network. *arXiv:2505.18806*.

Yan, Z., Han, Y., & Jiang, Z. (2025). Generative Adversarial Networks for Cybersecurity: A Comprehensive Review of Defense and Attack Applications. *arXiv:2509.20411*.

Zhang, Q., Liu, J., & Li, Y. (2024). A systematic review of deep learning techniques for enhancing network security through intrusion detection. *arXiv:2402.17020*.

Zhou, M., & Yang, B. (2025). *A deep learning-based framework for adaptive intrusion detection systems*. *arXiv:2505.05810*.

Polemi, N., Praça, I., Kioskli, K., & Bécue, A. (2024). Challenges and efforts in managing AI trustworthiness risks: a state of knowledge. Frontiers in big Data, 7, 1381163.

Munikoti, S., Agarwal, D., Das, L., Halappanavar, M., & Natarajan, B. (2023). Challenges and opportunities in deep reinforcement learning with graph neural networks: A comprehensive review of algorithms and applications. IEEE transactions on neural networks and learning systems, 35(11), 15051-15071.

Santoso, A., & Surya, Y. (2024). Maximizing decision efficiency with edge-based AI systems: advanced strategies for real-time processing, scalability, and autonomous intelligence in distributed environments. Quarterly Journal of Emerging Technologies and Innovations, 9(2), 104-132.