Role of AI in Cybersecurity

Aizada Kurmanalieva, Christy Peter, Mary Akrasi, Krishna Kagitha, Mohammed Sadequddin
Department of Computer Science and Information Systems, Elmhurst University, Illinois

akurm9685@365.elmhurst.edu, cpete9152@365.elmhurst.edu, kkagi6648@365.elmhurst.edu, makra5952@elmhurst.edu, m6760@elmhurst.edu

Abstract*— Artificial Intelligence is redefining modern cybersecurity by enabling intelligent, adaptive, and scalable defense mechanisms capable of countering today's rapidly evolving threat landscape. As organizations generate massive volumes of digital data across cloud platforms, IoT ecosystems, and hybrid networks, traditional signature-based security systems are no longer sufficient for detecting stealthy, automated, or previously unseen attacks. This paper provides a comprehensive analysis of AI-driven cybersecurity, examining key applications such as anomaly detection, intrusion detection systems, endpoint defense, user and entity behavior analytics, and automated incident response. In addition, the study evaluates foundational machine learning paradigms, advanced deep learning architectures, and reinforcement learning–based decision systems that enhance real-time monitoring and autonomous threat mitigation. Emerging challenges—including adversarial attacks, data imbalance, model drift, and explainability —are also explored, along with future trends such as generative AI–driven proactive defense, compound AI security architectures, digital-twin deception systems, and AI-augmented red teaming for quantum-resistant cryptography. By integrating technical insights, research gaps, and practical considerations, this paper highlights AI's transformative role in strengthening cyber resilience and shaping next-generation security ecosystems.

Index Terms - Artificial intelligence, anomaly detection, autoencoders, behavioral analytics, classification algorithms, clustering algorithms, convolutional neural networks, cybersecurity, data mining, data privacy, decision making, decision trees, deep learning, feature extraction, genetic algorithms, intrusion detection, intrusion detection systems, machine learning, machine learning algorithms, malware, malware detection, memory dumping, obfuscation, principal component

I. Introduction

Modern cybersecurity has evolved into one of the most essential disciplines in digital protection as societies increasingly depend on cloud systems, digital transactions, remote communication and interconnected devices across every sector of life.

This widespread digital transformation has dramatically expanded the volume of information being processed daily, creating new vulnerabilities and opportunities for exploitation by cybercriminals. As organizations migrate their operations to cloud platforms and rely heavily on online services, the threat landscape has multiplied in both size and complexity for economic governmental and operational activities.

Harvard Extension School 2024 emphasizes that today's attacks are no longer isolated or simplistic. They are highly coordinated, global in scale, and increasingly powered by automation and artificial intelligence. This means traditional cybersecurity approaches that are dependent on human supervision and rule-based detection

are no longer sufficient. The speed and sophistication of modern cyberattacks require systems that can operate at machine speed, adapt dynamically, and detect patterns that humans cannot identify.
As a result, Artificial Intelligence (AI) has become a transformative force within the cybersecurity world, reshaping how organizations anticipate, identify, and respond to emerging threats.explains that modern attacks are highly coordinated global in scope and often powered

by automation requiring defense strategies that exceed traditional security capabilities.

Cybersecurity now requires intelligent, adaptive systems capable of analyzing massive amounts of data and identifying unpredictable threats in real time. This shift has led to the widespread integration of AI within modern security frameworks, redefining both the capabilities and expectations of cybersecurity defense.

A. Basics of Cybersecurity

Cybersecurity is the practice of protecting digital systems networks and data from unauthorized access misuse or destruction. It encompasses the technologies, processes, and practices designed to protect digital systems, networks, applications, and data from unauthorized access or malicious damage.

As organizations of all sizes shift toward cloud infrastructure, digital workflows, and remote operations, cybersecurity has become a strategic priority across the world. Without effective protection, organizations risk data breaches, financial losses, service interruptions, legal penalties, and the erosion of user trust.

Traditional cybersecurity relied on signature-based antivirus tools, firewalls, intrusion detection systems, and predefined rules written manually by experts. While these methods were effective when threats were relatively predictable, they fail to address today's advanced attack techniques. Now, cybercriminals rely on:

- Zero-day vulnerabilities
- AI-generated phishing emails
- Ransomware-as-a-service kits
- Botnet-driven large-scale attacks
- Advanced Persistent Threats (APTs)
- Credential stuffing and automated brute-force attacks

Harvard Extension (2024) notes that attackers increasingly automate their operations, allowing them to target thousands of systems simultaneously, operate across borders, and evade basic forms of detection. These evolving threats demand cybersecurity systems that can learn, adapt, and respond much faster than human analysts. Modern cybersecurity therefore involves:

- Behavioral Analytics
- Real-time Threat Intelligence
- Automated Monitoring
- Data-driven risk assessment
- Continuous authentication
- Cloud and endpoint protection at scale

To meet these demands, AI-powered cybersecurity solutions have become indispensable.

B. Basics of Artificial Intelligence

Artificial Intelligence is the simulation of human intelligence in machines allowing them to reason, learn and make decisions. Artificial Intelligence enables machines to perform tasks that typically require human intelligence, including reasoning, learning, decision-making, and pattern recognition.
AI spans multiple subfields:

Machine Learning(ML): Algorithms learn from historical data to identify patterns.

Deep Learning: Neural networks analyze complex structures such as behavior sequences.

Natural Language Processing (NLP): Systems understand or mimic human language.

Autonomous Systems/ Agentic AI: AI systems plan, reason, and act independently.

Domo (2024) explains that *agentic AI, the* most advanced form of AI, goes beyond predictive analytics. These systems are capable of self-direction, autonomous decision-making, and strategic thinking. This represents a shift from AI models that only classify or predict data to AI systems that can dynamically respond to unfamiliar situations.

In cybersecurity, this means AI can:

- Identify suspicious patterns
- Investigate abnormal behavior
- Decide on containment actions
- Adjust its strategies in response to evolving threats
- Operate continuously without fatigue
- Detect complex intrusions that evade human analysis

This makes AI perfectly suited for the speed and complexity of cybersecurity analysis.

C. Why AI is needed in Modern Cybersecurity

AI is Crucial for Cyber Defense AI is indispensable in modern cyber defense for several reasons:

1. Cyber Threat Volume Exceeds Human Capacity: Large enterprises receive millions of security alerts per day. Excelsior University (2023) reports that human analysts cannot manually review or prioritize these alerts. AI rapidly processes large datasets and identifies genuine threats hidden within noise.

2. Attackers are Using AI: *AI-generated phishing emails, deepfake-based scams, malware that mutates automatically, and automated password attacks demonstrate the offensive use of AI. Defenders must use AI to match this increasing sophistication.*

3. AI Enables Early Detection Through Behavioral Analysis: *Instead of relying solely on known threat signatures, AI examines behavioral patterns. Examples include:*

Instead of relying solely on known threat signatures, AI examines behavioral patterns. Examples include:

- unusual login times
- abnormal file access
- atypical network requests
- unexpected data transfers

Harvard Extension (2024) explains that this method catches insider threats and stealthy attacks long before damage occurs.

4. AI Reduces False Positives

Traditional systems overwhelm analysts with irrelevant alerts. AI decreases false positives by identifying meaningful patterns and accurately classifying anomalies.

5. AI Enables Automated Incident Response

AI can autonomously:

- isolate infected devices
- stop suspicious sessions
- quarantine files
- disable compromised user accounts

This speed is essential during ransomware outbreaks where seconds matter.

6. AI Supports Scalability

- AI can manage thousands of systems simultaneously, whereas human analysts cannot.
- AI makes cybersecurity more proactive, efficient, and accurate than ever before.

KEY GLOBAL CYBERSECURITY STATISTICS AND TRENDS

Cybersecurity statistics indicate a rapidly worsening threat environment.

Excelsior University (2023) predicts global cybercrime costs will exceed $10 trillion annually by 2025 surpassing the cost of natural disasters and illegal drug markets combined.

Harvard Extension (2024) emphasizes a 150% increase in ransomware attacks, many involving double extortion and targeting government, healthcare, and education sectors. Attackers now use AI to write phishing messages that are:

- grammatically correct
- personalized
- context-aware

This greatly increases their success rate.The shift to remote work has expanded cyber exposure because:

- more home networks are insecure
- more devices connect to corporate systems
- cloud authentication is targeted heavily

Tools such as:

- Darktrace
- CrowdStrike Falcon
- IBM Watson Security

The tools mentioned above analyze billions of signals per second and outperform traditional systems significantly. These trends demonstrate why AI is central to modern cybersecurity strategies.

EVOLUTION OF AI IN CYBERSECURITY

AI's evolution in cybersecurity can be summarized in five stages:

Stage 1: Signature-Based Detection

- Dependent on known malware signatures

- Fails against new or modified threats

Stage 2: Machine Learning Models

- Detect anomalies by comparing activity to historical patterns
- Improved zero-day recognition

Stage 3: Deep Learning

- Uses neural networks
- Analyzes complex behaviors and large datasets
- Effective in intrusion detection, malware classification, fraud detection

Stage 4: Behavioral Analytics

- Creates baseline user profiles
- Identifies deviations
- Detects insider threats and credential compromise

Stage 5: Agentic / Autonomous AI

(Domo, 2024)

Agentic AI:

- reasons independently
- learns continuously
- adapts to ambiguous threats
- decides and acts without human intervention

This represents the future of hyper-responsive cyber defense and establishes the foundational understanding needed to explore the broader role of AI in cybersecurity. It highlights how cybersecurity has evolved from simple perimeter defense to a complex, data-intensive discipline requiring intelligent technologies. AI provides the necessary capabilities to analyze vast datasets, predict emerging threats, automate responses, and adapt to evolving attack strategies.

The global rise of cybercrime, the sophistication of attacks, and the expansion of organizational attack surfaces underscore the urgency of integrating AI into cybersecurity. Trends from Harvard Extension School (2024), Excelsior University (2023), and Domo (2024) confirm that AI is no longer optional but essential for resilience, accuracy, and advanced threat mitigation.

Together, these insights set the stage for deeper discussions on AI applications, machine learning methods, benefits, limitations, ethical concerns, and future trends.

II. AI Applications In Cybersecurity

With the rise of sophisticated cyber-attacks—including advanced persistent threats (APTs), zero-day exploits, polymorphic malware, and insider threats—traditional signature-based security mechanisms are no longer sufficient. Artificial Intelligence (AI), combined with Machine Learning (ML) and Deep Learning (DL), offers adaptive, data-driven detection and response capabilities. This paper explores critical AI applications in cybersecurity, including threat and anomaly detection, network and endpoint intrusion detection, user/entity behavior analytics (UEBA), and automated response orchestration. We examine real-world deployments, performance considerations, and open research challenges, illustrating the transformative impact of AI on cybersecurity operations.

I.Introduction: Role of AI and ML in Cybersecurity

Modern IT infrastructures—including cloud platforms, IoT devices, remote endpoints, and hybrid networks—generate massive volumes of heterogeneous data, such as logs, telemetry, and network flows. Traditional rule- or signature-based security tools struggle to keep up with emerging threats that exploit unknown vulnerabilities or mimic legitimate behavior. AI and ML offer effective solutions by learning baseline behavior and detecting deviations, enabling the identification of novel, stealthy, or polymorphic attacks that bypass conventional defenses [4], [5].
Intrusion detection remains a highly active research area within AI-cybersecurity integration [6]. Given alert overload, limited human resources, and rapidly evolving threats, AI-driven automation has become increasingly essential for modern security operations.

II. Threat Detection and Anomaly Detection

A.Importance of Anomaly-Based Detection

Traditional signature-based detection identifies only known threats with pre-defined patterns. In contrast, AI and ML models can establish a baseline of normal behavior and flag deviations as potential threats. This approach is particularly effective for zero-day exploits, polymorphic malware, and other previously unseen

attacks [5], [7]. Studies show that ML-driven anomaly detection increases coverage and adapts dynamically to new attack techniques [5].

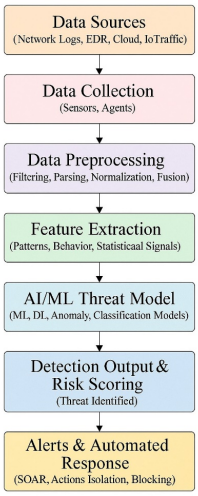

Fig. 1 AI Threat Detection Pipeline

B. Detection Pipeline Architecture

1. Data Collection: Network flows, packet metadata, system logs, endpoint telemetry, and authentication records.

2. Preprocessing & Feature Extraction: Cleaning, normalization, and extraction of features, e.g., session durations, packet statistics, login patterns.

3. Model Training/Baseline Modeling:
- *Unsupervised models* (clustering, auto-encoders) to define normal behavior.
- *Supervised or semi-supervised models* when labeled data is available [8].

4. Real-Time Monitoring & Inference: Score live data against baseline models.

5. Anomaly Detection & Risk Scoring: Identify deviations exceeding thresholds.

6. Alert Generation & Response: Analysts review alerts or trigger automated actions.

This pipeline allows organizations to detect unknown threats, including insider attacks and stealthy intrusions, without constant signature updates.

C. Use Cases and Effectiveness

AI-driven threat detection excels in the following scenarios:
- Detection of zero-day exploits and polymorphic malware.
- Monitoring large-scale network traffic in cloud or IoT environments.
- Identifying abnormal user behavior, such as unusual login times or data exfiltration.
- Scaling detection capabilities without linear increases in human analysts.

Empirical studies show ML-based systems detect a broader spectrum of threats with higher recall than traditional IDS/IPS tools [7], [9].

III. Intrusion Detection Systems (IDS) and Endpoint Detection & Response (EDR)

A.AI-Enhanced Network & Host-based IDS

AI-enhanced Network IDS (NIDS) can analyze packet flows, traffic behavior, and connection patterns beyond simple signature matching. This allows detection of stealthy intrusions, DDoS attacks, lateral movements, and data exfiltration [10]. AI-driven IDS adapts to evolving network patterns and reduces false positives compared to static rule-based systems [11], [12].

B. Behavioral Detection at Endpoint Level (EDR)

AI-powered Endpoint Detection and Response (EDR) monitors detailed endpoint telemetry, including process creation, network connections, file activity, system calls, and behavioral sequences. This enables early detection of file-less malware, ransomware, privilege escalation, and insider threats [12]. AI-based EDR provides continuous behavioral monitoring, threat scoring, and automated containment, representing a significant shift in proactive endpoint defense [13].

C. Integrated Network and Endpoint Defense Strategy

Layered deployment of both network-level and endpoint-level AI systems enhances visibility, correlates anomalies across domains, and strengthens defense against multi-stage attacks.

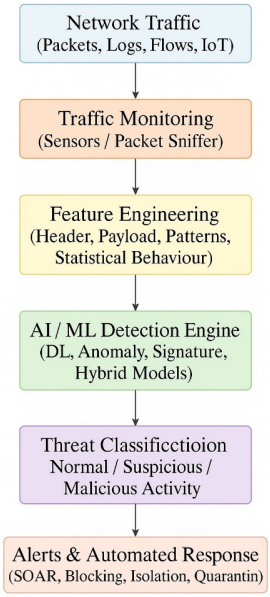

Fig. 2 Integrated IDS/EDR Architecture

IV. Behavioral Analytics and UEBA (User/Entity Behavior Analytics)

A. Concept and Value

UEBA extends anomaly detection to human users and non-network entities (e.g., servers, IoT devices, cloud accounts). By modeling behavioral profiles—login times, access patterns, resource usage, and inter-system interactions—UEBA can detect insider threats, compromised credentials, account takeover, and lateral movement without identifiable malware [14], [15].

B. Typical UEBA Workflow

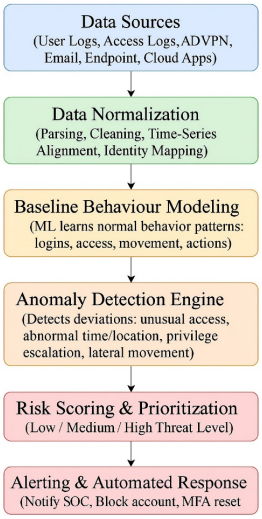

Fig. 3 UEBA Workflow Model:

1. Data ingestion from authentication logs, endpoint telemetry, network logs, and application services.
2. Baseline creation per entity (user, device, account) based on historical behavior.
3. Continuous monitoring and feature extraction (login frequency, data transfers, access patterns, location, time-of-day).
4. Anomaly scoring and risk classification; deviations flagged for review.
5. Alerting and response; automated actions triggered based on policy.

C. Use Cases and Advantages

- Early detection of account takeover and compromised credentials.
- Identification of insider threats or unauthorized data exfiltration.
- Monitoring user behavior across cloud, remote endpoints, and hybrid infrastructure.
- Reducing reliance on static permission or role-based access control systems.

V. Automated Incident Response and AI-Driven Security Automation

A.Response Workflow and Automation Pipeline

AI combined with Security Orchestration, Automation, and Response (SOAR) enables rapid mitigation, containment, and remediation of threats [16].

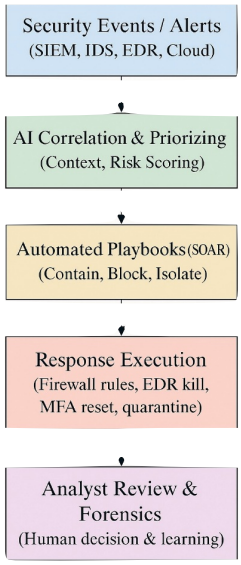

Fig. 4 Automated Incident Response Flowchart (placeholder).

Key Steps:

- Threat/anomaly detection and risk scoring by AI.
- Decision engine determines manual review or automatic response.
- Automated actions: isolate endpoints, block suspicious traffic, revoke credentials, or trigger sandboxing/forensic capture.
- Logging and feedback for auditing and model retraining.

B. Benefits

- Reduced Mean Time to Detect (MTTD) and Mean Time to Respond (MTTR).
- Immediate containment of fast-moving threats (e.g., ransomware).
- Reduced workload for security operations teams.
- Consistent, policy-driven response reduces human error.

C. Challenges

- False positives require careful tuning and human validation for sensitive systems.
- Privacy and compliance issues arise from behavioral and endpoint monitoring.
- Continuous monitoring and retraining are necessary due to model drift.
- Explain ability is critical for analyst trust and effective response.

VI. Research Insights, Limitations, and Open Challenges

AI in cybersecurity still faces several key challenges:

- Scalability: Processing real-time telemetry from large networks demands efficient models and optimized pipelines [14].
- Data quality: ML models require comprehensive, representative datasets [5], [14].
- Adversarial attacks: Attackers may attempt to evade detection or poison training data [14], [15].
- Privacy and compliance: Behavioral analytics involve sensitive data; compliance with GDPR/ CCPA is essential [14].
- Explainability: Deep learning models often lack interpretability; XAI techniques are needed [14], [15].
- Automation vs. human control: Fully automated response can be risky; hybrid human-AI oversight is recommended [14].

VII. Future Trends and Directions

- Autonomous ML / AutoML IDS frameworks: Reduce human effort in data processing, feature engineering, and model tuning [16].
- Cross-domain and federated learning: Share threat intelligence while preserving privacy and ownership.
- Explainable AI (XAI): Build trust and simplify audits in cybersecurity.
- Behavioral analytics in cloud, IoT, and hybrid environments: Essential for distributed infrastructures.
- Integration with threat intelligence and automated response: Enable pre-emptive defense and rapid mitigation.

AI-powered cybersecurity represents a major advancement from reactive, signature-based defense to proactive, intelligent, and adaptive protection. Instead of relying solely on predefined rules or known attack signatures, modern AI systems continuously learn from network behavior, enabling early detection of subtle anomalies and emerging threats. Through techniques such as anomaly detection, AI-enhanced IDS/EDR platforms, and behavioral analytics, organizations gain deeper visibility into system activities and can identify attacks that would otherwise evade traditional defenses.

Anomaly detection models establish baselines for normal user and network behavior, flagging deviations such as unusual account access patterns, abnormal data transfers, or suspicious process executions. Likewise, AI-driven IDS/EDR solutions analyze endpoint behavior, system calls, and traffic flows to detect malware, lateral movement, ransomware, and fileless attacks in real time. User and entity behavior analytics (UEBA) adds another layer, helping identify insider threats, compromised accounts, and privilege misuse by examining long-term behavioral trends.

Automated incident response further enhances security operations by enabling rapid containment actions, such as isolating compromised devices or blocking malicious traffic, often without human intervention. Although challenges remain—such as privacy concerns, resource demands, adversarial evasion, and the need for explainable AI—ongoing research continues to improve the accuracy, transparency, and robustness of AI-driven defenses. As threats grow more automated and sophisticated, AI-enabled security is positioned to become a central pillar of future cybersecurity infrastructure.

III. Machine Learning Techniques Used

*A. Foundational ML P*aradigms

The rapid proliferation of digital infrastructure and sophisticated attack vectors has elevated Machine Learning (ML), a core component of Artificial Intelligence (AI), into an indispensable tool for securing modern computational environments. ML is projected to positively transform the field of cybersecurity by enabling the discovery of critical data insights and automating complex defensive tasks that are overwhelming for human analysts[17]. Given the constant generation of massive amounts of data—often referred to as streaming data—current Intrusion Detection Systems (IDS) are continually tested to their limits[18]. ML algorithms are applied directly to this generated data to recognize patterns indicative of malicious activity, thereby enhancing the efficacy of network monitoring software and devices[18]. The effective implementation of ML in cyber defense requires a nuanced understanding of the three principal learning paradigms: supervised, unsupervised, and reinforcement learning. The choice of which approach to use for an intrusion detection scheme is inherently challenging, as each solution possesses its own distinct set of advantages and drawbacks[17].

1) Supervised Learning (SL): High-Velocity Classification of Known Threats

Supervised Learning (SL) represents the most common application of ML in cybersecurity, focusing on classification tasks using extensive, pre-labeled datasets. This paradigm trains models to establish a definitive mapping between input features and known threat classes. SL is the dominant method for recognizing established attack signatures, including robust spam filtering, phishing detection, and the classification of known malware families [19].

Successful SL deployment mandates a rigorous sequence of steps, beginning with data preprocessing, followed by effective feature selection, the application of classification techniques, and culminating in evaluation using appropriate metrics. Key algorithms utilized in this domain include:

- Decision Trees (DT) and Random Forests: Ensemble methods providing high accuracy and interpretability, frequently applied in email and network classification.
- Naive Bayes (NB): A simple, probabilistic classifier favored for spam filtering due to its speed and high classification accuracy, demonstrating results up to 99.46% in some comparative studies[19].
- Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN): Used across various applications, including intrusion detection and the analysis of phishing URLs.

The operational effectiveness of SL is, however, inherently constrained by its dependency on labeled data. Since security threats, such as polymorphic malware and zero-day exploits, constantly adapt and evolve, the SL model can only reliably classify threats for which it has been explicitly trained. This structural limitation means that SL is fundamentally inadequate as a sole defense mechanism against novel, unseen attack vectors, necessitating the integration of other paradigms.

2) Unsupervised Learning (UL): Discovery of Anomalies and Zero-Day Threats

Unsupervised Learning (UL) offers a unique advantage by operating without the requirement of pre-defined labels. In cyber defense, UL algorithms define a baseline of 'normal' system or network behavior and subsequently identify deviations or anomalies from this baseline. This capability makes UL exceptionally well-suited for uncovering zero-day threats—attacks exploiting previously unknown vulnerabilities that possess no existing threat signatures.

Algorithms commonly employed in UL anomaly detection include:

K-Means Clustering: This method groups similar benign data points into dense clusters. Data points that fall far outside these established clusters are flagged as potential anomalies.

Autoencoders (AE): These are Deep Learning models used to learn a compact, low-dimensional representation of normal data. When malicious or novel data is processed, the model fails to reconstruct the input

accurately, resulting in a large reconstruction error that signals an anomaly [20].

While UL provides a powerful mechanism for detecting novel threats, its operational deployment is hindered by a critical trade-off. The primary objective is to maximize coverage (Recall) for rare, critical events. To ensure such coverage, the model must be highly sensitive. However, increased sensitivity inevitably leads to non-malicious but unusual events being flagged as threats, resulting in a high rate of False Positives (FPs). This phenomenon, often termed the Anomaly Detection Paradox, creates alert fatigue for security operations center (SOC) personnel, potentially causing them to disregard legitimate high-severity alerts. Consequently, robust UL deployments require sophisticated filtering mechanisms or post-processing stages to mitigate the operational burden while retaining the essential coverage needed for zero-day threats[20].

3) Reinforcement Learning (RL): Adaptive Policy Optimization and Autonomous Response

Reinforcement Learning (RL) approaches the challenge of cyber defense not as a classification problem, but as a control problem. RL agents learn optimal, adaptive defensive policies by interacting dynamically with the network environment, making a sequence of decisions to maximize a long-term goal, such as minimizing the damage or duration of a breach [21]. This paradigm is driving the automation of complex cyber defense tasks, shifting capabilities from passive detection toward active incident response.

Deep Reinforcement Learning (DRL) architectures, such as Deep Q-Networks (DQN) and adapted models like A3C-Security (an asynchronous advantage actor–critic model), are utilized for security optimization. Advanced frameworks, such as the Adaptive RL Framework for Automated Cybersecurity Incident Response Strategy Optimization (ARCS), illustrate the sophistication of this approach [22]. The ARCS methodology employs a hierarchical state representation module to process raw security events and capture complex temporal dependencies. This state is then fed into a dual-stream architecture:

1. Immediate Stream (Q-Network): This component focuses on rapid, tactical responses, estimating the Q-value for each potential action based on the current state. This determines the most effective immediate action (A) to an event.

2. Long-Term Stream (Policy Network): This component is dedicated to strategic planning, learning a long-term policy ($\pi(s_t)$) designed to minimize the overall impact of the incident over an extended period, thereby guiding sustained defense actions.

RL uniquely frames defense as a series of actions over time, managing both instantaneous threats and long-term system integrity, a capability beyond traditional classification models. However, autonomous defense systems must overcome issues of trust and accountability. To address this, the integration of Explainable RL (XRL) provides security teams with clear justifications for machine-driven defensive actions, clarifying how the agent interprets network states and how its actions relate to long-term defensive goals.

B. Deep Learning Architectures for Advanced Threat Analysis

Deep Learning (DL) models, a subset of ML, are critical for extracting robust, non-linear features directly from raw data, bypassing the fragility of manual feature engineering. The application of DL methods, specifically Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), for network analysis and intrusion detection is a primary focus of current research [17].

1) Convolutional Neural Networks (CNNs): Extraction of Structural and Spatial Features

CNNs are highly effective at capturing spatial patterns in input data by applying convolutional filters. In cybersecurity, this principle is adapted to treat raw binary files, system call sequences, or network traffic data akin to images [23].

CNNs excel at extracting hierarchical spatial features from complex data types. In malware detection, this involves identifying localized patterns within raw binary or opcode sequences, such as structural anomalies often characteristic of obfuscated malware samples. The process typically involves applying 1D convolutional filters over the input sequence to extract local byte-level patterns. Subsequent max pooling layers are used to reduce dimensionality and introduce translational invariance, which significantly improves the model's capacity for generalization [23].

LEARNING PARADIGM	CYBERSECURITY TASK	TYPICAL ALGORITHMS	KEY OPERATIONAL CONSTRAINT
Supervised Learning (SL)	Intrusion/Malware Classification, Spam/Phishing Detection	SVM, Random Forest, Naive Bayes	Requires extensive, accurately labeled datasets for known threats.
Unsupervised Learning (UL)	Anomaly Detection, Zero-Day Threat Identification	K-Means Clustering, Autoencoders	Prone to high False Positive rates, leading to alert fatigue.
Reinforcement Learning (RL)	Automated Incident Response, Adaptive Policy Optimization	DQN, A3C, HRDL	Requires robust simulation environments and high trust/explainability (XRL).
Deep Learning (DL)	Advanced Feature Extraction, Malware/Traffic Analysis	CNN, RNN/LSTM, Hybrid Architectures	High computational resource intensity; eliminates manual feature engineering.

Table 1: Summary of Core ML Paradigms in Cybersecurity [21]

2) Recurrent Neural Networks (RNNs) and LSTMs: Modeling Sequential and Temporal Dependencies

Recurrent Neural Networks (RNNs) are architecturally distinct, designed specifically to process sequential information. They maintain a hidden state that acts as a memory, making them adept at learning long-range dependencies and temporal dynamics. RNNs are used to model sequences of communications between computers on a network to identify outlier network traffic.

For complex network traffic and malware behavior analysis, standard RNNs are often replaced by more advanced architectures like Long Short-Term Memory (LSTM) cells. LSTMs are effective at capturing complex relationships and long temporal patterns in network communications, overcoming the issue of vanishing gradients inherent in standard RNNs. This is crucial for tracing instruction flows and sequential behavioral signatures embedded within malware execution patterns. [24]

3) Hybrid Architectures: Synergistic Malware Detection

The most advanced applications of DL integrate CNNs and RNNs into hybrid architectures to achieve synergistic feature extraction. This design capitalizes on the strengths of both models: the CNN component extracts robust structural features (e.g., from raw bytes), while the RNN component (e.g., LSTM) analyzes the temporal or sequential relationships among those features (e.g., the sequence of system calls or network events).

This architectural fusion provides an end-to-end learning framework that bypasses the need for manual feature engineering, allowing the system to autonomously discover optimal feature representations directly from the raw data. This capability is essential when fighting polymorphic threats, which constantly mutate their code structure to evade traditional signature detection. By analyzing both the raw byte sequences (CNN) and the behavioral instruction flow (RNN), DL models secure high reliability against constantly evolving threats [24].

Experimental evaluations confirm the superior performance of these hybrid models. Studies involving malware detection show that a hybrid CNN-RNN model achieved an accuracy of 98.6% and a remarkably high recall rate of 98.9% on benchmark datasets, significantly outperforming classical methods like SVM and Random Forest. High recall is non-negotiable in security applications, as failing to detect even a single malicious file can lead to catastrophic consequences [24].

C. Data Benchmarking & Rigorous Performance Verification

1) Benchmark Datasets for Cyber Defense Research

Modern datasets are designed to simulate real-world threat complexity, moving beyond simplistic captures to include full network contexts and low-level system events.

NSD-KDD

The NSL-KDD dataset represents a crucial improvement over the earlier KDD'99 dataset, which suffered from inherent biases. The NSL-KDD refinement addresses two major deficiencies: the removal of redundant records in the training set and the elimination of duplicate records in the test set. This mitigation of data bias ensures that classifiers are not skewed towards more frequent records. Furthermore, by inversely proportional sampling from different difficulty levels, the NSL-KDD dataset achieves evaluation results that are more consistent and comparable across different research works[25].

CIC-IDS2017

This Intrusion Detection System (IDS) evaluation dataset is celebrated for resembling true real-world data, providing complete network traffic captures (PCAPs). The dataset was generated using a complete network topology, including firewalls, routers, and a variety of operating systems (Windows, Ubuntu, Mac OS X). It features labeled traffic across five days, encompassing normal activity and a comprehensive range of modern attacks. The availability of raw network flow data and extracted network and transport layer features provides the necessary fidelity to rigorously evaluate intrusion detection systems[26].

CIC-MalMem-2022

Specialized datasets are necessary to target sophisticated, obfuscated threats. The CIC-MalMem-2022 dataset focuses on memory dumping analysis, exposing general patterns and behaviors of obfuscated privacy malware at runtime. This dataset is balanced, consisting of 50% malicious and 50% benign memory dumps, totaling 58,596 records. It uses prevalent real-world malware families—including Spyware (e.g., Gator, TIBS), Ransomware (e.g., MAZE, Conti), and Trojan Horse (e.g., Zeus, Emotet)—and captures memory dumps in debug mode to simulate a realistic scenario where the dumping process does not interfere with the analysis[27].

The reliance on modern, low-level data sources, such as memory dumps and complete PCAPs, structurally confirms that high-level statistical network flow features are increasingly insufficient for threat detection. This trend validates the necessity of Deep Learning models that can process these high-dimensional, complex raw inputs directly.

2) Critical Evaluation Metrics for Imbalanced Classification

In cybersecurity, the vast discrepancy between benign data (majority class) and attack data (minority class) means that simply Accuracy is insufficient and often misleading as an evaluation metric. A model predicting all traffic as benign could achieve 99.9% accuracy but would be functionally useless. Therefore, domain-specific classification metrics are critical for assessing performance[28].

Precision (P): Measures the quality of positive predictions, specifically minimizing False Positives (FPs). High precision is essential to prevent alert fatigue among SOC personnel by reducing the noise level.[29]

Recall (R) or Sensitivity: Measures the completeness of capturing actual positive cases, maximizing True Positives (TPs). High recall is a top priority in high-stakes security applications where missing a single threat (False Negative) can be catastrophic[29].

F1 Score: The F1 Score is the harmonic mean of precision and recall. It is a robust measure of accuracy that balances both quality and (P) and completeness (R), making it highly effective for assessing performance in imbalanced datasets[28].

3) Advanced Threshold-Agnostic Metrics: ROC-AUC and PR-AUC

To provide a comprehensive performance measure independent of a specific classification threshold, advanced metrics are employed:

ROC-AUC (Receiver Operating Characteristic - Area Under the Curve): The ROC curve plots the True Positive Rate (Recall) against the False Positive Rate (FPR) across various threshold settings. The AUC quantifies the overall ability of the model to distinguish between positive and negative classes. ROC-AUC is useful in imbalanced datasets as it provides an aggregated performance measure[28].

PR-AUC (Precision-Recall Area Under the Curve): The PR curve plots Precision against Recall at different threshold settings. PR-AUC is the preferred metric for highly imbalanced classification models in security. This

preference stems from the fact that PR-AUC focuses exclusively on the performance concerning the positive (attack) class, offering a more stringent and realistic evaluation than ROC-AUC when the minority class is the primary operational concern[28].

The priority placed on Precision versus Recall, and consequently the preference for the F1 Score or PR-AUC, ultimately reflects the organization's tolerance for risk and operational workload. A security system protecting critical national infrastructure must prioritize high Recall, even at the cost of some FPs. Conversely, a high-volume enterprise must maintain high Precision and F1 scores to manage operational bandwidth.

4) Cybersecurity ML Model Workflow: Datasets to Evaluation

The application of ML in cybersecurity follows a structured, iterative workflow, often visualized as a pipeline extending from data acquisition to continuous system adaptation. This workflow highlights the necessity of feedback loops for adapting to the evolving threat landscape.

PHASE	STEP/ACTION	DATA/OUTPUT	KEY CONSIDERATION/CITATION
I. Data Acquisition	Source selection, traffic/system logging, data capture.	Raw PCAPs, Memory Dumps, System Logs	Requires high-fidelity, real-world representative data (e.g., CIC datasets).
II. Data Preprocessing & Engineering	Cleaning, normalization, feature extraction (for SL/UL), or raw data encoding (for DL).	Labeled Feature Vectors or Encoded Raw Data	NSL-KDD refinement addresses data bias and redundancy.
III. Model Selection & Training	Choose appropriate paradigm (SL, UL, RL, DL) based on threat context (Known vs. Novel).	Trained Model Weights/Policy	Deployment of complex architectures (e.g., hybrid CNN-RNN) for advanced feature learning.
IV. Model Evaluation & Optimization	Performance validation using the test set.	Confusion Matrix, Accuracy, P, R, F1, ROC-AUC, PR-AUC	Strict adherence to imbalance-appropriate metrics (PR-AUC) is required.
V. Deployment & Monitoring	Integration into operational IDS, endpoint detection, or SOAR systems.	Real-time Decisions/Actions	Requires continuous tuning to minimize False Positives and avoid alert fatigue.
VI. Feedback & Adaptation	Analyze prediction errors; collect new threat samples; update training regimen.	Updated Data, Retrained Model Policy	Crucial for adapting to evolving threats and mitigating evasion techniques.

Table 2. Cybersecurity ML Model Workflow: Datasets to Evaluation[26]

D. Synthesis and Future Directions

The investigation into Machine Learning applications in cybersecurity reveals that effective cyber defense requires a convergence of specialized learning paradigms. No single ML approach is sufficient to handle the diverse and dynamic nature of modern threats.

A truly resilient cyber defense system must employ Supervised Learning for rapid, high-confidence identification of known malware, phishing, and intrusion signatures. This must be complemented by Unsupervised Learning to provide crucial coverage against zero-day exploits and novel attack vectors, though careful tuning is essential to manage the resulting operational load associated with false positives. Furthermore, the capacity for Reinforcement Learning is necessary to enable autonomous, adaptive incident response, transforming the security system from a passive alert mechanism into an active defense automation platform (SOAR integration).

Critically, the foundation of all these paradigms is Deep Learning—specifically the hybrid use of CNNs for structural analysis and RNNs/LSTMs for sequential dependency modeling. DL enables the system to handle the complexity of raw network traffic and memory dumps, automatically discovering the underlying malicious logic necessary to defeat sophisticated techniques like polymorphism.

Future research efforts must continue to address key challenges to maximize the utility of these techniques. These include mitigating the high computational resource intensity required by advanced Deep Learning models, solving the persistent dilemma of high false positive rates in unsupervised anomaly detection, and enhancing the transparency of autonomous systems through Explainable AI (XAI) and Explainable Reinforcement Learning (XRL). Only through the comprehensive, coordinated deployment of these advanced ML and DL paradigms can cybersecurity keep pace with the evolving adversarial landscape[21].

IV. Benefits & Limitations

BENEFITS

Although the advantages of using AI in cybersecurity are widely recognized, researchers still point out an important gap in the existing studies. Many academic papers discuss how AI can improve security, but they often do not evaluate how well different AI techniques actually perform specific cybersecurity tasks. Methods such as machine learning (ML), deep learning (DL), and natural language processing (NLP) are being adopted more frequently, but most studies tend to look at them in a general way. Instead of comparing different algorithms for specific use cases, the literature usually focuses on broad applications.

In practice, cybersecurity tasks like intrusion detection, malware analysis, phishing detection, and anomaly identification each require different capabilities. For example, detecting malware might rely heavily on identifying patterns in code, while phishing detection often involves analysing text content or user behaviour. Despite these clear differences, very few studies examine which AI models are best suited for each type of task. This lack of detailed comparison makes it difficult for security professionals to understand which technique works best in real-world scenarios.

Sommer and Paxson (2010) also highlight a similar issue. They argue that many researchers apply machine learning techniques without fully evaluating whether these models are appropriate for real-world cybersecurity environments. According to them, it is not enough for an AI model to work well in a controlled or experimental setting. Instead, it must be tested for scalability, meaning its ability to handle large volumes of data; interpretability, meaning how easily its decisions can be understood; and adaptability, meaning its ability to respond to new or evolving threats. These practical factors are often overlooked, which results in AI models that might appear effective in academic research but struggle when faced with real-time, unpredictable cyberattacks.

In simple terms, although many ML and AI techniques show great potential, there is still limited understanding of how they behave in real operational environments—especially when dealing with new attack patterns or reducing false alarms. This gap in research creates an opportunity for further investigation.

For this reason, the present paper aims to provide a broad and structured review of different AI techniques used in cybersecurity and how they are applied. While it does not explore every individual technique in extreme detail, it offers a wide overview of AI models and their roles within the cybersecurity landscape. The goal is to give readers a clearer picture of what is currently known, what is still missing, and where future research can be directed. [30]

LIMITATIONS

The proposed system also has a few limitations, mainly related to how the cameras connect to the network. Some pipelines are located in very remote areas where wireless internet access is weak or completely unavailable. In such cases, the system cannot function properly because the cameras depend on a stable connection to transmit live video.

One possible solution is to install fiber-optic cables to connect the cameras in these remote locations to the nearest area with reliable internet coverage. Another option is to use VSAT (Very Small Aperture Terminal) satellite technology, which allows cameras to connect via satellite instead of relying on traditional internet networks.

However, both of these solutions still share a major challenge: if the internet connection becomes unstable, slow, or goes offline, the cameras in that area will stop transmitting data. When this happens, the footage cannot be displayed on the AR (augmented reality) glasses used for monitoring. As a result, important sections of the pipeline may not be visible in real time, leading to reduced monitoring quality and potential safety risks [31].

FLAG CHALLENGES

NYU CTF Dataset is a scalable, open-source benchmark designed to evaluate the performance of LLMs in solving cybersecurity Capture the Flag (CTF) challenges. Compiled from popular CTF competitions, the dataset includes diverse tasks and metadata tailored for LLM testing and adaptive learning. It supports advanced function calling and external tool integration, enabling a fully automated evaluation system with enhanced workflows. The dataset facilitates the assessment of five LLMs, encompassing black-box and open-source models, and compares their performance with human participants in interactive cybersecurity tasks. This benchmark provides a robust platform for advancing LLM capabilities in vulnerability detection, task automation, and real-world threat management.

The Dynamic Intelligence Assessment (DIA) framework introduces an innovative approach to evaluating AI models by leveraging dynamic question templates and advanced metrics to address the limitations of static benchmarks. The accompanying dataset, DIA-Bench, spans various disciplines, including mathematics,

cryptography, cybersecurity, and computer science, featuring diverse challenge formats such as text, PDFs, visual puzzles, and CTF-style tasks. By incorporating four novel metrics, DIA highlights gaps in model reliability and confidence, revealing frequent errors even with seemingly simple questions when presented in varied forms. Evaluations of 25 leading LLMs demonstrated challenges with complex tasks and unexpected inconsistencies in confidence levels, setting a new benchmark for assessing adaptive intelligence and self-awareness in AI systems. [33]

AI MISUSE OR THREATS

1) PASSWORD GUESSING

Password guessing is growing increasingly common, and the ability to guess is always improving. Under such situations, password security is under unprecedented demand. Password guessing has been approached from a variety of angles, including heuristic search, probabilistic models, and deep learning (DL).

The password guessing attack is categorized into offline password guessing and online password guessing based on whether the attack procedure requires interaction with the server or not. The former approach necessitates the authentication server storing the user account password file, after which the attacker guesses the password on the local host. The number of guesses that can be attempted in this situation is solely limited by the attacker ' s computational resources. The latter simply needs the attacker to be connected to the network and does not need a password file.

The number of guesses that can be attempted, however, is frequently limited by the server ' s security policy, such as the US National Identity Standards NIST-800-63-3, which states that the maximum number of failed logins allowed for a government website system in a month is 100, and the account will be locked if it exceeds 100. From a generative AI perspective one of the tools used is the PassGAN a combination of the term " Password " and Generative Adversarial Network. PassGAN employs the upgraded Wasserstein GAN to learn from the data distribution of billions of leaked passwords, with the goal of generating higher quality password guesses. D trains the G to replicate bogus passwords that are very near to the true password data distribution, attaining higher accuracy than HashCat and John the Ripper. [31]

2) RANSOMWARE

Ransomware attacks continue to be one of the most serious threats that enterprises and governments face. It has been established that criminals with very little training in information technology (IT) may be able to carry out complex ransomware attacks using chatbots driven by generative AI. Additionally, criminals with extensive IT knowledge but little other talents may be able to use generative AI to create more convincing phishing emails. In general, an average ransomware attack in the Web 2.0 era could result in a significant one-time ransom payment. In the metaverse, however, a ransomware attack might result in crypto jacking, in which ransomware takes over a user ' s device indefinitely and uses it to mine for cryptocurrencies in the background. A ransomware attack comprises six stages: (1) creating the malware; (2) deployment; (3) installation; (4) command and control; (5) destruction; and (6) extortion. [31]

3) SOCIAL ENGINEERING

Social engineering is one of the biggest challenges facing network security because it exploits the natural human tendency to trust. AI tools provide hackers with powerful capabilities to launch sophisticated social engineering attacks, including:

Sophisticated Personalization : AI-powered tools enable hackers to collect massive volumes of data from a variety of sources, including social media, public databases, and leaked data. Hackers can use this amount of information to create highly tailored spear phishing emails that appear legitimate and trustworthy. These tailored attacks have a far higher success rate and pose a major risk to individuals and businesses.
Deepfake Threats: Artificial intelligence algorithms may create realistic synthetic media, such as modified audio and video information. Deepfake technology is leveraged upon by hackers to impersonate trustworthy individuals, creating fraudulent content that tricks employees into exposing sensitive information or engaging in destructive behaviour. Because of their genuineness, deepfakes are much more difficult to detect. [31]

Automation and Scale: AI-powered technologies enable cybercriminals to automate many parts of the attack process, including as reconnaissance, email production, and reaction analysis. This automation enables hackers to undertake large-scale attacks that target several individuals at the same time. [31]

Evading Detection: AI-powered technologies enable cybercriminals to automate many parts of the attack

process, including as reconnaissance, email production, and reaction analysis. This automation enables hackers to undertake large-scale attacks that target several individuals at the same time. [31]

4)GENERATIVE AI INSPIRED ATTACKS

AI Inspired Cyber Threats In Metaverse A survey of these types of attack and these range from causative, exploratory, evasion, Whitebox, Blackbox, Gray box.

Causative attacks: In a causative attack, the adversary feeds the target classifier with training data after changing the labels of samples with DL scores that are far away from the decision boundary to reduce the reliability of the training process This type of attack incorporates both the poisoning attack and the backdoor attack. The poisoning attack reduces the model accuracy by adding malicious training data to the training process for the model.

Additional inaccuracy is provided through the backdoor attack which is an extended version of the poisoning attack and further trains the model on malicious training data with a trigger attached. In this stage, the model correctly recognizes the data without the trigger as usual, but the presence of certain triggers causes data to be misrecognized by the model. [31]

Exploratory attack: By changing the test data without gaining access to the training process, the exploratory attack causes the target model to misinterpret test data. The adversarial example is one form of exploratory attack. An adversarial example is a sample in which some noise has been added to the input data; to human eyes, it looks to be normal data, but the model misinterprets it. [31]

Evasion attack: This is the most common type of attack. During the testing phase, the adversary attempts to defeat the system by modifying damaging samples. This option is based on the assumption that the training data is unaffected. [31]

Whitebox attack: The Whitebox attack on a machine learning model, an adversary has complete knowledge about the model (for example, the type of neural network and the number of layers). The attacker is aware of the training algorithm (for example, gradient descent optimization) and has access to the training data distribution. He also understands the parameters of the entire trained model architecture. The adversary uses this information to examine the feature space in which the model may be vulnerable, i.e., where the model has a high error rate.

The model is then abused by altering an input with the adversarial example creation method. Both medical picture classification and segmentation tasks have achieved great success rates in white box attacks. Once the medical image recognition model and network structure parameters are fully protected, access to the model is restricted, making white box assaults with high attack success rates and black box attacks requiring a large number of queries difficult. White-box Attack obtains the parameters and structure of the model. White-box attacks have some gradient-based methods for specific models. [31]

Blackbox Attack: A Blackbox attack assumes no prior knowledge of the model and exploits it by utilizing information about the model' s settings and prior inputs. For example, in an oracle attack, the adversary explores a model by providing a series of carefully generated inputs and observing outputs. Adversarial learning is a severe threat to real-world machine learning applications. While various countermeasures exist, none of them can be a one-size-fits-all. [31]

CLUSTERING: PROS & CONS

Another approach that could be used to classify suspicious activities on user accounts is clustering. By grouping user accounts into homogeneous groups, based on the type of activity carried out (frequency of user posts, time spent on the platform, frequency of user logins, and more), it is also possible to identify suspicious activities that may concern multiple user accounts compromised by the same attacker, whose purpose could be, for example, to spread spam messages or publish unwanted posts by coordinating the activities of the various accounts.

Clustering is, in fact, an approach that allows the detection of similarities (even hidden ones) within various user groups; once grouped into different clusters, we will need to determine which of these clusters are actually representative of suspicious activity, and, within each cluster, which accounts are involved in possible fraudulent activity.

However, even in the case of clustering, it is necessary to carefully choose the type of algorithm to use: in fact, not all clustering algorithms are effective in detecting suspicious activity. For example, clustering algorithms, such as k-means, require the correct determination of the number of clusters (by defining, in advance, the value of the parameter k, from which the algorithm takes its name), a feature that is not very suitable for the detection of suspicious user activities in practice, because we are not usually able to define the correct number of clusters in which the accounts must be grouped. Furthermore, algorithms such as k-means do not work with features expressed in the form of categories or binary classification values. [32].

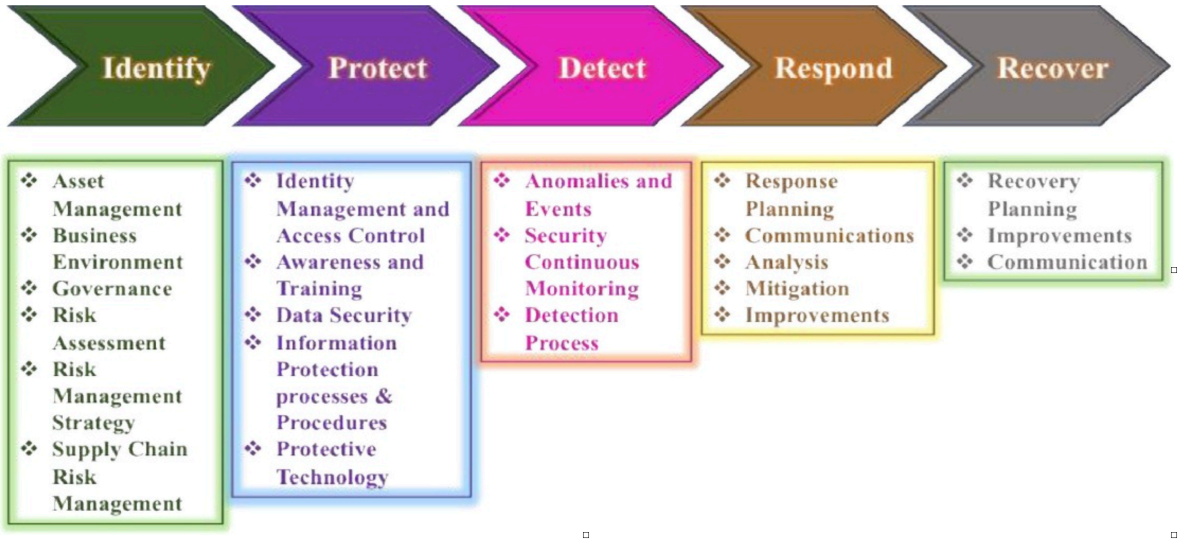

Fig 5. Cybersecurity Framework

ASPECT	PROS (BENEFITS)	CONS (LIMITATIONS/ETHICAL ISSUES)
Speed	Detects threats instantly	May produce false alerts quickly
Automation	Reduces manual workload	Overreliance may reduce human oversight
Accuracy	Learns from patterns to detect unknown threats	Accuracy depends on high-quality data
Scalability	Handles large volumes of data	Expensive to deploy on enterprise scale
Bias & Fairness	Can standardize security decisions	May inherit bias from training data
Privacy	Enhances monitoring for safety	Risks over-surveillance
Misuse Risks	Protects networks and users	Attackers can also weaponize AI

Table 3. Clustering: Pros and Cons

V. Future Trends in AI driven Cybersecurity

Artificial Intelligence (AI) is accelerating a paradigm shift in cybersecurity, enabling defense systems that are adaptive, predictive, and increasingly autonomous. As cyberattacks evolve into multi-vector, automated, and strategically deceptive operations, AI-powered security architectures are transitioning from reactive monitoring to proactive and self-optimizing protection. Building on findings from methodologies such as generative modelling for cyber-defense, explainable and dimensionally-reduced intrusion detection, LLM-driven domain threat analysis, digital-twin–based deception for cyber-physical resilience, and AI-augmented red teaming for quantum-resistant cryptography, this section outlines the emerging trends expected to define the next decade of cyber defense.

A. Generative AI driven proactive defense ecosystems

Generative AI (GAI) will transform cybersecurity from a reactive paradigm to a proactive threat anticipation model. Unlike traditional pattern-based systems, which rely on historical data, generative architectures can create synthetic attack scenarios, simulate phishing campaigns, generate malware variations, and model adversarial behavior before attackers deploy them in real environments. This capability fundamentally strengthens early-warning systems and reduces the time to detect and mitigate emerging threats.

1) AI-Generated Synthetic Threat Environments

GAI models—such as LLMs, diffusion models, and transformer-like adversarial generators—can produce synthetic malware, zero-day variants, and extensive phishing datasets. These models enable defenders to pre-train intrusion detection models using realistic threat samples, including polymorphic and obfuscated attack chains. This accelerates model robustness, even in data-scarce environments. Figure 6 provides the conceptual overview of how generative pipelines integrate with cybersecurity training workflows, reinforcing the shift toward "threat synthesis as defense".[34]

2) GAI-Enhanced Automated Red-Teaming

Future cyber ranges will rely heavily on GAI-driven red teaming capable of autonomously generating:

- adversarial payloads,
- phishing lures tailored to specific organizations,
- privilege escalation exploits,
- social engineering scripts.

Attacker-side automation will force defenders to adopt equivalent automation[34]. Cybersecurity will evolve into an "AI-vs-AI" environment, where defenders must continuously simulate attacks in order to harden systems.

3) Anticipatory Defense and Predictive Analytics

Through generative modeling, defenders can simulate entire kill chains, forecast vulnerability exploit likelihood, and generate synthetic network telemetry. As shown in (Fig. 7 and Fig. 8), probabilistic threat modeling will become a fundamental element of SOC operations.[34]

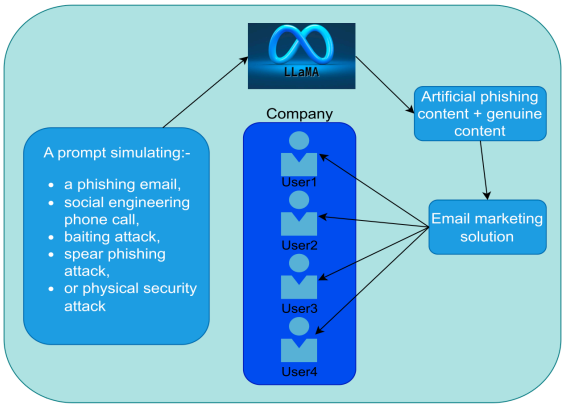

Fig. 7 Phishing resilience training using Meta LLaMA.

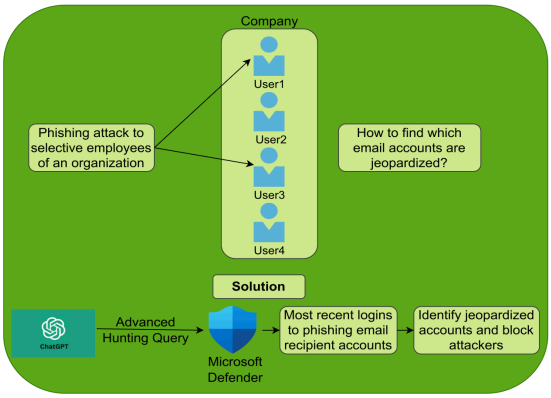

Fig. 6 Security Code generation using Open AI's ChatGPT

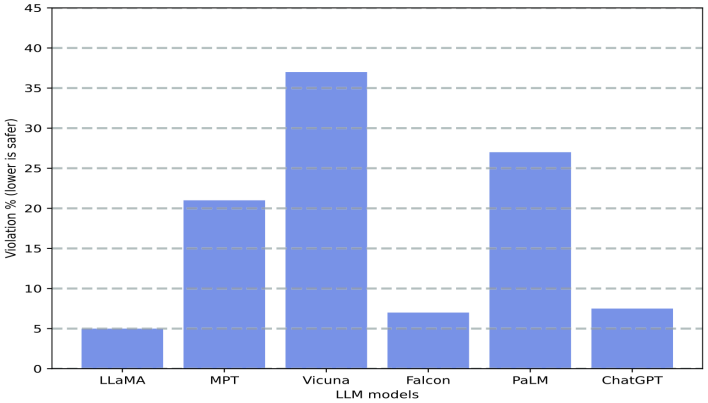

Fig. 8 Safety human assessment outcomes for LLaMA 2-Chat in comparison to other models - both open and closed-source.

These figures illustrate how GAI integrates with security workflows, enabling dynamic threat generation and iterative learning loops.

B. Compound AI Systems and Multi-Layered Detection Architectures

A defining trend for AI-driven cybersecurity is the rise of compound AI architectures—systems built from multiple specialized components (LLMs, vector databases, retrieval systems, rule-engines, and validators). These systems outperform traditional ML because they can detect novel, hybrid, and semantic attack patterns.

1) Emergence of Compound AI Pipelines

DomainLynx [36] is a leading example of a compound AI system, combining:

- vector embeddings (OpenAI 1536-dimensional model),
- Weaviate vector DB,
- HNSW indexing for approximate nearest neighbors,
- LLM-based threat scoring,
- a Threat Recognition Validation (TRV) module

Such architectures represent the future of detection pipelines: scalable, modular, and capable of analyzing millions of events daily with semantic reasoning.

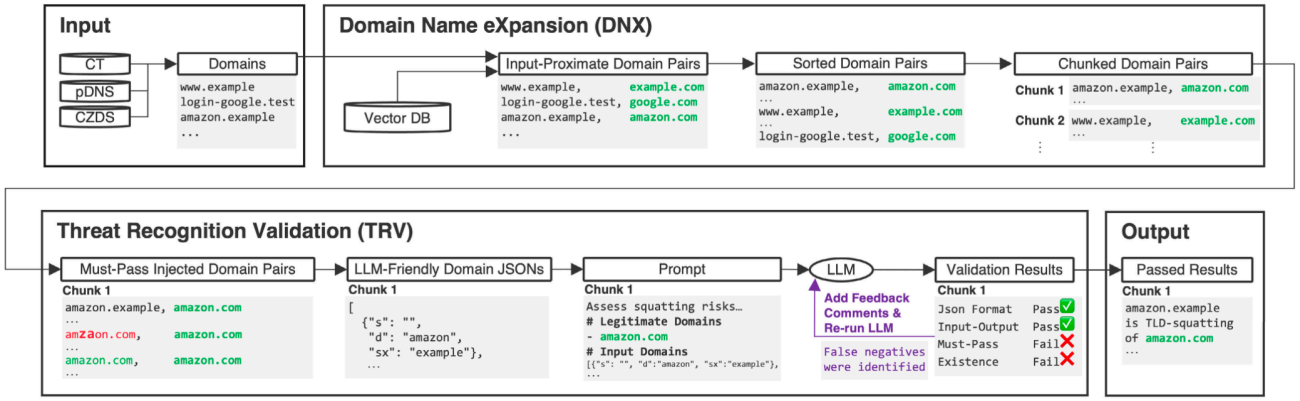

Fig. 9 DomainLynx system architecture overview.

Article 3, Figure 1 shows how DomainLynx's system architecture integrates multi-stage AI components to detect domain squatting tactics, including hybrid and combination squatting, with 94.7% accuracy across large-scale datasets [36].

2) Hybrid Threat Detection

Future systems must detect compound threats combining typo-squatting, homographs, level-squatting, and semantic impersonation. DomainLynx shows that traditional rule-based systems fail against such hybrid threats; LLM-based systems excel due to their semantic understanding of linguistic and visual patterns.

3) Scale-Adaptivity and Real-Time Operations

DomainLynx processed 2.09 million domain registrations in one month, identifying 34,359 suspicious domains—a scale impossible for human analysts. Future SOCs will adopt similar pipelines to monitor CT-logs, pDNS feeds, WHOIS data, and DNSSEC anomalies at global scale. [36]

C. Explainable AI (XAI) as a Mandatory Requirement in Future Security Systems

As AI becomes increasingly integrated into cybersecurity operations, explainability is emerging as a mandatory requirement rather than an optional feature. Deep learning models often behave like "black boxes," making it difficult for analysts to understand why specific alerts or classifications are generated. This lack of transparency poses challenges in high-stakes environments where accountability, regulatory compliance, and audit-ability are essential. Security teams must be able to justify automated decisions, especially when actions such as

account lockouts, access restrictions, or system isolation are triggered by AI-generated outputs.

Explainable AI addresses these issues by providing insights into feature importance, model reasoning, and decision pathways. By revealing how an AI model derives its conclusions, XAI increases trust, supports faster incident validation, and helps analysts detect potential biases or anomalies in model behavior. It also plays a crucial role in identifying vulnerabilities to adversarial attacks and monitoring model drift over time [35].

1) XAI-Integrated Deep Learning Models

The necessity for XAI within intrusion detection systems (IDS) is quite high. The EIDCDR-XAIADL model integrates:

- deep learning–based detection,
- class decomposition,
- feature relevance profiling,
- explainability layers using SHAP/LIME.

Fig. 10 illustrates the complete workflow of an explainable IDS pipeline, showing how interpretability augments traditional model outputs.

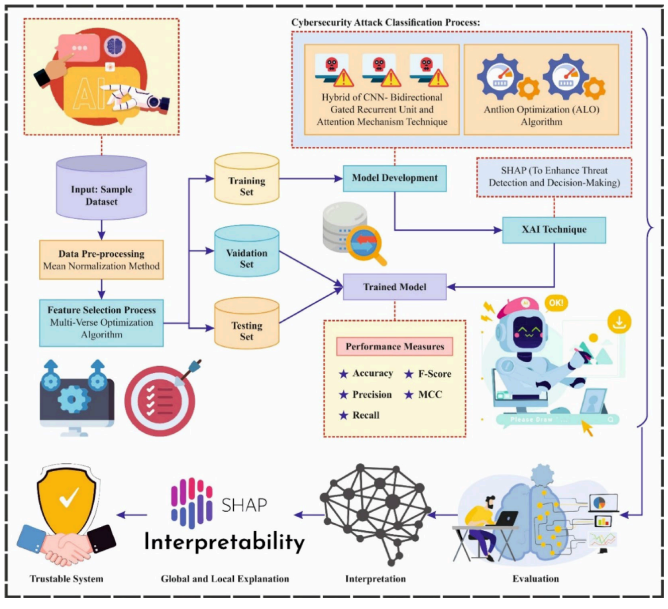

Fig. 10 Overall workflow of EIDCDR-XAIADL model

2) Regulatory and Compliance Pressures

Explainability will be required for compliance with:

- EU AI Act,
- U.S. NIST AI Risk Management Framework,
- critical infrastructure regulations.

Security teams must justify automated decisions (e.g., account lockouts, anomaly flags). Without XAI, ML-based security products will be rejected in high-assurance environments.

3) XAI for Adversarial Attack Resistance

Explainable systems can detect when attackers manipulate input features to fool AI models. Transparency helps uncover poisoning attacks, drift anomalies, and model tampering—making XAI a core part of AI security hardening.[35]

D. Autonomous SOCs and AI-Driven Incident Response

Cybersecurity operations will increasingly be handled by autonomous or semi-autonomous AI agents capable of:

- alert triage,
- root-cause analysis,
- incident containment,
- remediation execution,
- report drafting.

1) AI-Enhanced Operational Efficiency

AI reduces SOC workload by automating:

- log analysis,
- correlation of multi-source telemetry,
- anomaly scoring.[35]

Future SOCs will use:

- AI copilots for analysts,
- automated policy recommendations,
- AI-generated firewall or SIEM rules,
- real-time autonomous threat hunting.

2) Adaptive Learning SOC Environments

Next-generation SOCs will continuously self-train using:

- synthetic attack data from GAI,
- active feedback loops,
- reinforcement learning (RL) agents.

This creates a living SOC ecosystem capable of automatically adjusting to novel threats.

E. Advanced LLM-Driven Phishing, Domain, and Social Engineering Detection

1) Semantic Phishing Detection

LLMs will enable anti-phishing systems that analyze:

- linguistic sentiment,
- semantic meaning,
- context consistency,
- style-based impersonation.

This allows detection of sophisticated phishing content generated by GAI.

2) Large-Scale Domain Abuse Detection

DomainLynx [36] proves LLMs can outperform baseline systems in detecting hybrid squatting domains, using embeddings rather than simple string matching.

3) Multimodal LLM-Powered Detection

Future systems will combine:

- webpage screenshots,
- email headers,
- DNS metadata,
- certificate fingerprints,
- conversational tone analysis.

This multimodal AI model will outperform single-modality detection.

F. AI-Driven Deceptive Cyber-Resilience for Critical Infrastructure

A major future direction in AI-based cybersecurity is the rise of proactive deception systems that mislead attackers rather than simply detecting them. Article 4 introduces a next-generation paradigm where digital twins act as active traps, creating synthetic PV generation states to strategically divert adversaries from real assets [37]. Instead of static rule-based defenses, AI-generated digital twin environments evolve continuously, presenting attackers with high-fidelity but fabricated operational data. As attackers unknowingly target these digital replicas, real systems remain insulated, creating a shift from reactive defense to attacker-engagement strategies.

Reinforcement learning emerges as a second critical component: deep RL agents autonomously learn optimal mitigation policies under dynamic attack conditions [37]. This is essential for future smart grids where adversaries leverage automation and AI themselves. RL enables cyber-defense models to anticipate adversarial actions, adapt strategies in real time, and coordinate deception patterns across multiple digital twins. Furthermore, the combination of deception, RL, and game-theoretic attacker modeling establishes a multi-layered adaptive defense ecosystem that will shape AI-driven security in future renewable energy infrastructures [37].

Blockchain-integrated quantum-secured authentication adds another emerging trend. Article 4 demonstrates how combining blockchain with quantum-resistant verification safeguards PV control operations from unauthorized tampering while supporting real-time dispatch efficiency [37]. As cyber-physical systems become highly distributed, decentralized authentication and post-quantum protection will become indispensable.

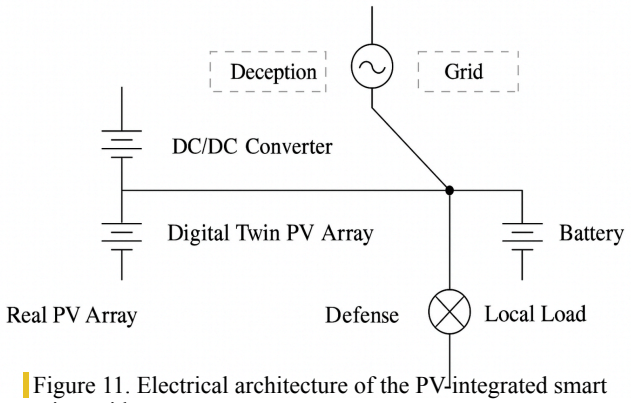

Figure 11. Electrical architecture of the PV-integrated smart microgrid.

G. AI-Enhanced Red Teaming for Quantum-Resistant Cryptography

A second transformative trend is AI-integrated penetration testing for quantum-era security. Classical cryptography is becoming inadequate against quantum computing threats, requiring new modes of continuous adversarial simulation. The study introduces an AI-driven red teaming framework that probes vulnerabilities in BB84 quantum key distribution (QKD) and NIST-approved post-quantum algorithms [38]. These AI models generate adversarial probes, perform protocol fuzzing, simulate side-channel attacks, and reveal flaws that traditional formal validation cannot detect.

Machine learning systems — including transformer architectures, generative adversarial networks, and NLP-driven models — create semantic misuse patterns and dynamic exploit payloads to stress-test quantum-resistant protocols under realistic attack conditions [38]. This represents a fundamental shift: AI is no longer just a defensive tool but a critical component in validating next-generation cryptography. By generating high-volume, high-complexity attack traces, AI enables quantum security standards to undergo continuous refinement before deployment.

The red teaming framework also integrates continuous feedback loops, where each AI-identified vulnerability leads to protocol hardening, followed by re-testing under new adversarial conditions [38]. This iterative cycle represents the future of quantum-era security assurance—continuous, adaptive, and AI-augmented.

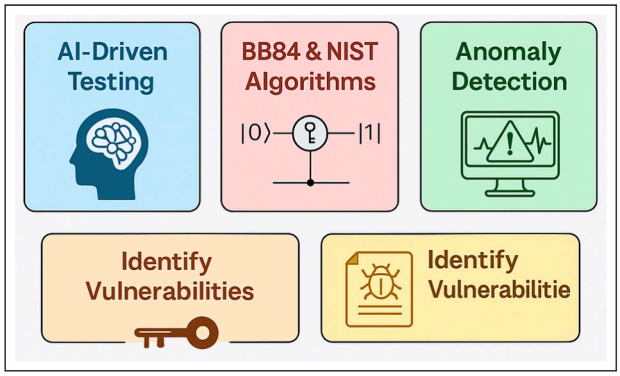

Fig. 12 Red teaming quantum-resistant cryptographic standards: framework integrating artificial intelligence and quantum security.

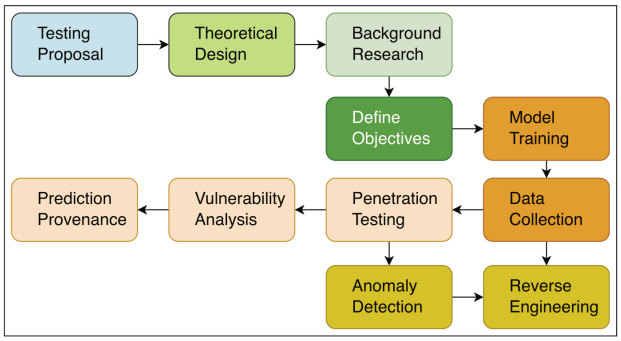

Fig. 13. Penetration testing approaches for the BB84 quantum cryptography protocol.

VI. Conclusion

Artificial Intelligence has become an indispensable component of modern cybersecurity, enabling defense systems that are more adaptive, proactive, and resilient than traditional rule-based security approaches. Through machine learning, deep learning, and reinforcement learning, AI enhances the detection of zero-day attacks, automates incident response, and provides scalable monitoring across complex digital environments. Advanced methods, including hybrid CNN-RNN architectures, behavioral analytics, and explainable AI frameworks, significantly improve accuracy while reducing false alarms—addressing long-standing challenges in intrusion detection and threat analysis. However, the integration of AI also introduces new risks, such as adversarial manipulation, model uncertainty, and the growing misuse of generative AI for sophisticated attacks. As cyber threats become increasingly automated and intelligent, the future of cybersecurity will depend on compound AI systems, autonomous SOC operations, digital-twin deception technologies, and AI-enabled validation of quantum-resistant cryptographic standards. Overall, this research demonstrates that AI is not merely an enhancement to cybersecurity but a foundational driver of next-generation defense, requiring continuous innovation, transparency, and strategic governance to ensure secure and trustworthy digital ecosystems.

VII. References

[1] Domo. (2024). *Agentic AI explained: Definition, benefits, and use cases*. https://www.domo.com/blog/agentic-ai-explained-definition-benefits-and-use-cases

[2] Excelsior University. (2023). *The role of artificial intelligence (AI) in cybersecurity.* https://www.excelsior.edu/article/ai-in-cybersecurity/

[3] Harvard Extension School. (2024). *AI and the future of cybersecurity.* https://extension.harvard.edu/blog/ai-and-the-future-of-cybersecurity/

[4] N. Mohamed, "Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms," Journal Name, 2025.

[5] M. Paramesha, N. L. Rane, J. Rane, "AI, Machine Learning and Deep Learning for Cybersecurity Solutions: Emerging Technologies and Applications," PU MRJ, vol. 1, issue 2, 2024.

[6] K. Achuthan et al., "Advancing cybersecurity and privacy with artificial intelligence," Journal / Conference, 2024.

[7] P. R. Brandao et al., "Exploring the Role of Artificial Intelligence in Detecting Advanced Persistent Threats," MDPI journal, 2025.

[8] M. Rahman, S. Al-Shakil, M. R. Mustakim, "A survey on intrusion detection system in IoT networks," Computer & Security Journal, 2025.

[9] J. A. Carter, E. R. Thompson, M. D. Perez, S. L. Hastings, "AI-Based Endpoint Detection and Response: Design and Evaluation," Cybersecurity Research, 2023.

[10] S. Otoum, B. Kantarci, H. Mouftah, "A Comparative Study of AI-based Intrusion Detection Techniques in Critical Infrastructures," arXiv preprint, 2020.

[11] L. Yang, A. Shami, "Towards Autonomous Cybersecurity: An Intelligent AutoML Framework for Autonomous Intrusion Detection," arXiv, 2024.

[12] "AI for Cybersecurity: Threat Detection & Automated Incident Response," Payoda Report, 2025.

[13] "Endpoint Detection and Response (EDR): The complete guide," Vectra AI white-paper, 2025.

[14] "User and Entity Behavior Analytics (UEBA) Overview," Cybersecurity Glossary, 2025.

[15] "Network Behavior Anomaly Detection (NBAD)," Security Reference Article, 2019.

[16] "AI in Cybersecurity & Automated Incident Response Solutions," Industry Report, 2025.

[17] Alshuaibi, A., Almaayah, M., & Ali, A. (2025). Machine Learning for Cybersecurity Issues : A systematic Review. Journal of Cyber Security and Risk Auditing, 2025(1), 36–46.

[18] B. I. Seraphim, S. Palit, K. Srivastava and E. Poovammal, "A Survey on Machine Learning Techniques in Network Intrusion Detection System," *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, 2018, pp. 1-5, doi: 10.1109/CCAA.2018.8777596.

[19] AlShaikh, M., Alrajeh, Y., Alamri, S., Melhem, S., & Abu-Khadrah, A. (2025). Supervised methods of machine learning for email classification: a literature survey. *Systems Science & Control Engineering*, *13*(1). https://doi.org/10.1080/21642583.2025.2474450

[20] S. Oluwadare and Z. ElSayed, "A Survey of Unsupervised Learning Algorithms for Zero-Day Attacks in Intrusion Detection Systems," *Proc. International FLAIRS Conference*, vol. 36, no. 1, 2023. doi: 10.32473/flairs.36.133182

[21] Abubakar, M. (2025). Explainable Reinforcement Learning for Adaptive Cyber Defense in Encrypted Networks. Preprints. https://doi.org/10.20944/preprints202511.0668.v1

[22] Ren, S., Jin, J., Niu, G., & Liu, Y. (2025). ARCS: Adaptive Reinforcement Learning Framework for Automated Cybersecurity Incident Response Strategy Optimization. *Applied Sciences*, *15*(2), 951. https://doi.org/10.3390/app15020951

[23] P. Avhad, S. Kolse, P. Pangavhane, T. Gadekar and A. Sangale, "Deep Learning for Malware Detection and Classification," 2023 4th International Conference on Computation, Automation and Knowledge Management (ICCAKM), Dubai, United Arab Emirates, 2023, pp. 1-6, doi: 10.1109/ICCAKM58659.2023.10449531.

[24] Shakhlo Dustova, Mehriniso Abuzalova, Nigora Adizova, Dilnoz Ruzieva, and Yarash Ruziev "AI-powered malware detection: a hybrid CNN-RNN model for real-time threat analysis", Proc. SPIE 13803, Optical and Computational Technologies for Measurements and Industrial Applications (OptiComp 2025), 138032P (22 September 2025); https://doi.org/10.1117/12.3078041

[25] M Hassan Zaib, NSL-KDD Network Security, Information Security, Cyber Security, https://www.kaggle.com/datasets/hassan06/nslkdd

[26] Intrusion Detection Dataset, https://www.unb.ca/cic/datasets/ids-2017.html

[27] D. Cevallos-Salas, F. Grijalva, J. Estrada-Jiménez, D. Benítez and R. Andrade, "Obfuscated Privacy Malware Classifiers Based on Memory Dumping Analysis," in IEEE Access, vol. 12, pp. 17481-17498, 2024, doi: 10.1109/ACCESS.2024.3358840.

[28] Y.Friedman , "Understanding F1 Score, Accuracy, ROC-AUC, and PR-AUC Metrics for Models"

[29] Becerra-Suarez, F.L.; Tuesta-Monteza, V.A.; Mejia-Cabrera, H.I.; Arcila-Diaz, J. Performance Evaluation of Deep Learning Models for Classifying Cybersecurity Attacks in IoT Networks. *Informatics* 2024, *11*, 32. https://doi.org/10.3390/informatics11020032

[30] Ofusori, L., Bokaba, T., & Mhlongo, S. (2024). *Artificial intelligence in cybersecurity: A comprehensive review and future direction*. Applied Artificial Intelligence, 38, e2439609.

[31] *Artificial Intelligence and Metaverse Through Data Engineering*, edited by Jagdish Chandra Patni, Nova Science Publishers, Incorporated, 2024. *ProQuest Ebook Central*.

[32] Parisi, Alessandro. Hands-On Artificial Intelligence for Cybersecurity : Implement Smart AI Systems for Preventing Cyber Attacks and Detecting Threats and Network Anomalies, Packt Publishing, Limited, 2019. ProQuest Ebook Central.

[33] Dynamic Intelligence Assessment: Benchmarking LLMs on the Road to AGI with a Focus on Model Confidence. N. Tihanyi, T. Bisztray, R.A. Dubniczky, R. Toth, B. Borsos, B. Cherif, R. Jain, L. Muzsai, M.A. Ferrag, R. Marinelli, et al. Proceedings of the 2024 IEEE International Conference on Big Data (BigData).

[34] S. Sai, U. Yashvardhan, V. Chamola and B. Sikdar, "Generative AI for Cyber Security: Analyzing the Potential of ChatGPT, DALL-E, and Other Models for Enhancing the Security Space," in *IEEE Access*, vol. 12, pp. 53497-53516, 2024, doi: 10.1109/ACCESS.2024.3385107.

[35] Alamro, H., Alahmari, S., Nemri, N., Aljebreen, M., Alhashmi, A. A., Alamro, S., Alqazzaz, A., & Al Duhayyim, M. (2025). Enhanced intrusion detection in cybersecurity through dimensionality reduction and explainable artificial intelligence. *Scientific Reports*, *15*(1), 1–25.

[36] Chiba, D., Nakano, H., & Koide, T. (2025). DomainLynx: Advancing LLM Techniques for Robust Domain Squatting Detection. *IEEE Access*, *13*, 29914–29931.

[37] Li, B., Jin, X., Ba, T., Pan, T., Wang, E., & Gu, Z. (2025). Deceptive Cyber-Resilience in PV Grids: Digital Twin-Assisted Optimization Against Cyber-Physical Attacks. *Energies (Basel)*, *18*(12), 3145.

[38] Radanliev, P. (2025). Red teaming quantum-resistant cryptographic standards: a penetration testing framework integrating AI and quantum security. *Journal of Defense Modeling and Simulation*.

About The Authors

Aizada Kurmanalieva is currently pursuing the Master of Science in Computer and Information Technology at Elmhurst University. She is originally from the city of Osh, Kyrgyzstan and completed her Bachelor's degree in World Economics from Kyrgyz-Russian Slavic University. Her academic interests include artificial intelligence, data engineering, cloud computing, and cybersecurity, with project experience in building data-driven applications and network management solutions. She aims to further contribute to the field through research and industry applications in advanced computing and intelligent systems.

Christy Kunjumon Peter is currently pursuing the Master of Science in Computer and Information Technology at Elmhurst University, Illinois, USA. He is originally from New Delhi, India, and completed his Bachelor of Technology in Computer Science at Amal Jyothi College of Engineering, Kanjirapally, Kerala, graduating in 2024. His academic and research interests include artificial intelligence, cybersecurity, and intelligent systems, with a specialized focus on evaluating AI models and developing guardrails for safe and reliable AI deployment. He aims to contribute to the advancement of secure, transparent, and responsible AI technologies through ongoing research and practical innovation.

Krishna Prasanna Kagitha is a graduate student in the M.S. in Computer and Information Technology program at Elmhurst University. She completed her B.Tech in Electronics and Communication Engineering from Jawaharlal Nehru Technological University, Kakinada (India). Her academic and professional interests include emerging technologies, artificial intelligence, cybersecurity, and data-driven systems. She is currently involved in work related to AI-based threat detection and aims to contribute to innovative research and real-world technological solutions in the future.

Mary Akrasi is a graduate student in the M.S. in Computer and Information Technology program. Originally from Accra, Ghana, she earned her Bachelor's degree in Political Science. Her research and professional interests span cybersecurity, artificial intelligence, cloud and network infrastructure, and data engineering, with recent work focusing on DNSSEC trust anchors, SQL-based data analysis, and enterprise network configuration projects. She aspires to pursue innovative research and practical solutions in emerging technologies, secure computing, and data-driven systems

Mohammed Sadequddin is currently pursuing the Master of Computer and Information Technology at Elmhurst University. He is originally from Hyderabad, India and completed his Bachelor's degree in Bachelors of Commerce from St. Joseph's Degree & PG College. His academic interests include areas such as artificial intelligence, cybersecurity, data analytics with project experience Role of AI in Cybersecurity. He aims to further contribute to the field through research and industry applications in advanced computing and intelligent systems.