

AI-Powered Cyber Attacks and Defence Mechanisms

Dr. Shilpa B. Sarvaiya¹, Meher Gopal Mankar², Pritam Ashok Borade³

Head, Department of MCA¹

Students, Department of MCA^{2,3}

Vidya Bharti Mahavidyalaya, Amravati, Maharashtra, India

sarvaiya.shilpa@gmail.com, meherpatil84@gmail.com, pritamborade24@gmail.com

Abstract: *Artificial Intelligence (AI) is becoming an important component of modern cybersecurity. Machine Learning (ML), Deep Learning (DL), Natural Language Processing (NLP), and Generative AI are increasingly used to analyse security data, identify malicious behaviour, detect malware, classify phishing attempts, and support incident response. At the same time, the same technologies can be misused by cybercriminals to improve phishing, social engineering, synthetic-media fraud, vulnerability discovery, malware development, and other attack activities. This dual-use character of AI has created a rapidly changing security environment. This paper presents a literature-based study of AI-powered cyberattacks and AI-based defence mechanisms, focusing mainly on research and threat reporting from 2020 to 2026. It examines AI-assisted attack techniques, defensive applications, adversarial machine learning, traditional versus AI-based security, real-world threat trends, limitations, privacy, ethics, and future research. The study finds that AI can improve the speed and scale of security analysis, but its effectiveness depends on data quality, continuous monitoring, secure model development, explainability, and human oversight. The paper argues that AI should complement rather than replace established controls such as authentication, encryption, patching, network segmentation, endpoint protection, and backups. A layered architecture combining conventional security, AI analytics, continuous evaluation, and human decision-making is therefore recommended.*

Keywords: Artificial Intelligence, Cybersecurity, Machine Learning, Deep Learning, Generative AI, Adversarial Machine Learning, Phishing, Malware Detection, Intrusion Detection, Threat Intelligence, Deepfake, Zero Trust

I. INTRODUCTION

Cybersecurity has become a fundamental requirement for organisations that depend on digital systems, cloud platforms, mobile devices, Internet of Things (IoT) devices, online banking, e-commerce, healthcare systems, industrial systems, and government services. The expansion of digital infrastructure has simultaneously increased the number of systems and identities that can be targeted by attackers. Modern attacks are rarely limited to a single technical action. They can involve reconnaissance, social engineering, initial access, credential theft, privilege escalation, persistence, lateral movement, data collection, exfiltration, and extortion.

Traditional cybersecurity has historically depended on firewalls, antivirus signatures, predefined rules, access-control mechanisms, encryption, vulnerability scanning, and manually maintained indicators of compromise. These mechanisms remain essential. However, security teams now process extremely large volumes of telemetry from endpoints, networks, cloud services, applications, identities, and user activity. The scale and speed of this information make manual inspection alone difficult.

Artificial Intelligence provides a complementary approach. ML models can identify statistical patterns in security data, classify events, detect unusual behaviour, and assist analysts in prioritising alerts. Deep learning can process complex



and high-dimensional information, while NLP can analyse emails, reports, web content, and other text. Generative AI can also assist analysts with summarisation, investigation, documentation, and response planning.

The same capabilities can be misused. Attackers can use AI to generate convincing communications, automate portions of reconnaissance, create synthetic content, analyse information, and adapt some attack activities. ENISA's 2025 threat landscape discusses the growing role of AI and large language models in phishing and social engineering [4]. Microsoft's 2025 Digital Defense Report similarly describes AI as both a defensive capability and an emerging component of the threat environment [5].

The result is a dual-use environment. Defenders can use AI to improve detection and response, while attackers can use related technologies to increase speed, scale, and realism. Cybersecurity research must therefore address both AI for security and the security of AI systems themselves.

A. Problem Statement

Traditional security controls may have difficulty identifying unknown threats, processing large volumes of events, and adapting quickly to changing attacker behaviour. AI-based security can address some of these problems, but it introduces its own risks, including data poisoning, adversarial inputs, model drift, privacy concerns, false positives, and explainability challenges. The central problem addressed by this research is how AI-powered attacks can be understood and countered through secure, reliable, and responsible defence mechanisms without abandoning proven conventional security controls.

B. Aim

The aim of this research is to examine how Artificial Intelligence is changing both cyberattack techniques and cybersecurity defence mechanisms, and to identify practical principles for using AI securely and responsibly.

C. Objectives

The objectives are: (1) to study major AI applications in cybersecurity; (2) to identify important AI-assisted cyberattack techniques; (3) to examine AI-based defence mechanisms; (4) to compare traditional and AI-based cybersecurity; (5) to study adversarial machine-learning risks; (6) to analyse ethical, privacy, and governance concerns; (7) to identify research gaps; and (8) to suggest future research directions and practical recommendations.

D. Research Questions

RQ1: How is AI being used to improve cyberattacks? RQ2: How is AI being used to improve cyber defence? RQ3: What limitations affect AI-based cybersecurity? RQ4: How can AI systems be protected from adversarial manipulation? RQ5: How should AI and conventional controls be combined in a practical security architecture?

II. RESEARCH METHODOLOGY

This study follows a literature-review and comparative-analysis methodology. It does not claim an original penetration-testing experiment or an independently measured detection benchmark. The evidence base consists of peer-reviewed research papers, conference publications, standards, government documents, cybersecurity reports, and recognised technical publications. The primary period considered is 2020–2026, reflecting the rapid development of machine learning, generative AI, deepfake technology, cloud security, and adversarial machine learning during this period.

Sources were identified from academic and professional repositories such as IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, Google Scholar, NIST, ENISA, Microsoft Security, and other established cybersecurity organisations. The selected literature was compared according to AI technique, security application, attack category, defence category, reported benefits, limitations, risks, and future research needs.

The study uses qualitative synthesis rather than a statistical meta-analysis. This approach is appropriate because the reviewed works use different datasets, evaluation metrics, threat models, and deployment environments. A direct numerical comparison between unrelated experiments could therefore be misleading. Instead, the paper identifies common findings and recurring challenges across the literature.



A. Inclusion Criteria

Priority was given to sources published between 2020 and 2026, authoritative government and standards publications, peer-reviewed literature, and recognised industry threat reports. Older foundational references were retained where they provide essential theoretical background. Sources that lacked identifiable publication information or verifiable bibliographic details were not treated as primary evidence.

B. Analysis Framework

Each source was examined under six broad dimensions: AI technology, attack or defence use case, operational benefit, limitation, security or ethical risk, and future research direction. This structure supports the comparative tables presented later in the paper.

III. LITERATURE REVIEW

Research on AI and cybersecurity has expanded significantly as organisations have adopted machine learning for intrusion detection, malware classification, anomaly detection, fraud detection, and threat intelligence. Ozkan-Okay et al. surveyed AI and ML techniques for cyber security and described the use of machine learning, deep learning, and reinforcement learning across multiple security tasks [3]. Their review illustrates the breadth of AI-based cybersecurity research but also highlights the need to consider dataset quality and operational limitations.

NIST's AI Risk Management Framework provides a broader perspective on trustworthy AI. It identifies characteristics including validity and reliability, safety, security and resilience, accountability and transparency, explainability and interpretability, privacy enhancement, and fairness [1]. These characteristics are directly relevant to security systems because a model that detects threats but cannot be trusted, explained, monitored, or protected may introduce additional organisational risk.

Adversarial machine learning has developed into a distinct research area. NIST's 2025 taxonomy describes attack categories including evasion, poisoning, privacy attacks, and misuse against predictive and generative AI systems [2]. This is important because an AI security system can itself become a target. A defender may deploy an accurate classifier, but an attacker who understands its weaknesses may manipulate inputs or training data to reduce its effectiveness.

Recent threat reports also demonstrate that AI is entering an already active cybercrime ecosystem. ENISA's 2025 Threat Landscape analysed thousands of incidents and discussed the use of AI in phishing and social engineering [4]. Microsoft's 2025 Digital Defense Report similarly describes changing patterns involving identity attacks, phishing, ransomware, and AI-enabled activity [5].

The literature therefore supports two connected research directions. The first is "AI for cybersecurity," where AI is used to detect and respond to threats. The second is "security of AI," where models, datasets, APIs, agents, and AI-enabled applications must themselves be protected. A complete cybersecurity strategy must address both.

A. Research Trend

A notable trend from 2020 onward is the movement from isolated machine-learning classifiers toward integrated security workflows. Instead of using AI only to label a network packet or file, modern security platforms increasingly combine telemetry collection, event correlation, risk scoring, analyst assistance, and response orchestration. Generative AI has further expanded this trend by supporting natural-language interaction with security data. However, increased automation also increases the importance of access control, auditability, and human approval for high-impact actions.

IV. AI-POWERED CYBERATTACKS

AI-Assisted Cyberattack Lifecycle

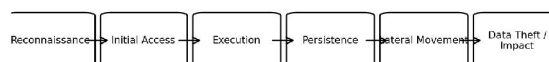


Fig. 1. AI-assisted cyberattack lifecycle.



AI-powered attacks should not be understood as completely new attack categories. In many cases, AI acts as an enabling technology that makes existing activities faster, more scalable, or more convincing. The following categories are especially relevant to current research.

A. AI-Assisted Phishing and Social Engineering

Phishing remains a major method of obtaining initial access. Conventional phishing messages often contain generic language, obvious grammatical errors, suspicious links, or inconsistent branding. Generative AI can reduce some of these weaknesses by helping attackers produce natural language, translate messages, personalise content, and generate many variations.

AI can potentially analyse publicly available information about a target and use that information to construct a more contextually appropriate message. This raises the difficulty of relying only on spelling mistakes or generic phishing indicators. Defence therefore needs multiple signals, including domain reputation, link analysis, identity controls, authentication, behavioural detection, and user awareness.

ENISA's 2025 threat landscape identifies phishing as an important intrusion vector and discusses the use of large language models to enhance social-engineering activities [4]. The significance is not that every phishing campaign is AI-generated, but that AI can lower the effort required to produce convincing content at scale.

B. Deepfakes and Synthetic Identity

Generative models can produce synthetic images, voices, and videos. In cybersecurity, this can enable impersonation and fraud. Examples include fake executive communications, voice impersonation, manipulated video meetings, synthetic identity material, and attempts to defeat weak identity-verification procedures.

The security implication is that biometric or audiovisual evidence should be considered in context rather than treated as automatically trustworthy. Strong authentication, independent verification channels, transaction controls, and contextual risk analysis can reduce the impact of synthetic-media attacks. Microsoft has highlighted the increasing security relevance of synthetic content and identity fraud in its recent threat reporting [5].

C. AI-Assisted Malware and Evasion

AI can assist attackers in analysing malicious software, generating variations, or identifying characteristics that may influence security detection. The more important issue is adaptability: attackers may use automation to experiment with different approaches more rapidly than manual processes allow.

Defensive measures should therefore combine static analysis, dynamic analysis, behavioural monitoring, endpoint telemetry, sandboxing, threat intelligence, and conventional signatures. A model should not be evaluated only against known malware samples; it should also be tested against changing and previously unseen behaviours.

D. Automated Vulnerability Discovery

AI can analyse source code, configuration files, application behaviour, documentation, and system relationships to identify potential weaknesses. This capability can benefit defenders because it can accelerate vulnerability management, but it can also be misused.

The most practical defensive response is not to avoid AI-based analysis. Instead, organisations should combine automated discovery with secure software development, patch management, asset inventories, configuration management, and validation by security professionals. Automated findings should be prioritised according to exploitability and business impact.

E. Adversarial Machine Learning

Adversarial ML focuses on attacks against machine-learning systems. Evasion attacks attempt to alter inputs so that a model produces an incorrect result. Poisoning attacks attempt to influence training data. Privacy attacks attempt to



extract information about training data or model behaviour. Misuse attacks involve abusing AI capabilities for unintended purposes.

NIST's 2025 adversarial ML taxonomy provides terminology and a structured view of these threats [2]. The key lesson is that AI models must be treated as security-sensitive components. Data provenance, access controls, secure training pipelines, adversarial testing, model monitoring, and incident-response plans should therefore be included in AI deployments.

F. Password and Credential Abuse

AI can support analysis of patterns in leaked information and can assist attackers in scaling social-engineering or credential-related activities. However, many successful credential attacks continue to depend on conventional weaknesses such as password reuse, stolen credentials, insufficient authentication, and exposed services. This reinforces the importance of multi-factor authentication, password managers, credential monitoring, least privilege, and strong identity governance.

G. AI-Assisted Reconnaissance

Reconnaissance involves collecting information about organisations, technologies, people, and exposed services. AI can help organise and summarise large quantities of information, potentially reducing the manual effort required for target profiling. From a defensive perspective, organisations should maintain accurate external asset inventories, minimise unnecessary exposure, monitor leaked credentials, and continuously review publicly exposed information.

V. AI-BASED CYBER DEFENCE MECHANISMS

AI-based defence mechanisms are designed to help security teams process large datasets, detect unusual behaviour, classify threats, prioritise alerts, and support response. Their value is highest when they are integrated with conventional controls and monitored continuously.

A. Anomaly Detection

Anomaly detection is one of the most common applications of AI in cybersecurity. A model can learn statistical characteristics of normal activity and identify deviations. Examples include unusual login locations, abnormal data-transfer volumes, unexpected application execution, atypical network communication, or sudden changes in user behaviour.

Unsupervised and semi-supervised techniques are particularly attractive where labelled attack data is limited. However, an anomaly is not necessarily an attack. Legitimate operational changes can also look unusual. Human review and contextual correlation are therefore essential.

B. Machine-Learning Intrusion Detection

Machine-learning intrusion detection systems classify network or host activity. Supervised models can learn from labelled traffic, while unsupervised methods can identify clusters and deviations without complete labelling. Deep learning can process complex patterns but may require substantial training data and computing resources.

Important evaluation measures include precision, recall, false-positive rate, false-negative rate, detection latency, and resource consumption. A high accuracy score alone may not indicate practical usefulness when a dataset is imbalanced or does not represent current traffic.

C. Malware Detection

AI can assist malware detection through both static and dynamic analysis. Static analysis examines properties such as file structure, imported libraries, metadata, and code characteristics. Dynamic analysis observes behaviour such as process creation, file modification, registry changes, network communication, and system calls.



A layered approach can combine signatures for known threats with behavioural models for unknown or modified samples. Such a design reduces dependence on any single technique.

D. Threat Intelligence and Security Analytics

Security teams process logs, vulnerability information, threat feeds, incident reports, and alerts from multiple systems. AI can correlate events and identify relationships that may not be obvious when sources are examined separately. NLP can extract entities such as malware families, vulnerabilities, threat actors, affected products, and indicators from unstructured reports.

The benefit is reduction of analyst workload. The limitation is that automatically generated correlations can contain errors, so security analysts should be able to inspect the evidence supporting an AI-generated conclusion.

E. Security Operations Centre Assistance

AI can support SOC analysts by grouping duplicate alerts, summarising incidents, recommending investigation steps, and prioritising high-risk events. Generative AI can also help convert technical logs into readable summaries. However, access to security tools should be tightly controlled. An AI assistant with permission to disable accounts, delete files, or change firewall rules could cause serious damage if manipulated or misconfigured.

A safer approach is graduated automation: low-risk repetitive actions can be automated, while actions with significant operational impact require human approval.

F. Behavioural Authentication

AI can analyse behavioural signals such as login timing, device characteristics, access patterns, and transaction behaviour. This supports risk-based authentication, in which multiple signals contribute to a security decision. The approach must address privacy and false positives. A legitimate user may behave differently while travelling, working remotely, or changing devices. Systems should therefore allow appropriate review and recovery procedures rather than treating every anomaly as malicious.

G. Fraud and Transaction Monitoring

Financial institutions and e-commerce systems can use AI to detect unusual transaction patterns. A model may consider transaction amount, frequency, device information, location, account history, and behavioural context. AI can help prioritise suspicious activity, but automated blocking should be designed carefully because false positives can inconvenience legitimate customers.

H. Cloud and IoT Security

Cloud and IoT environments generate large amounts of telemetry. AI can support configuration monitoring, anomaly detection, identity analysis, and device-behaviour profiling. IoT devices may have limited resources and long lifecycles, making lightweight detection and centralised monitoring particularly important. In cloud environments, identity and access management, configuration security, and continuous monitoring remain fundamental even when AI analytics are deployed.

TABLE I: COMPARISON OF TRADITIONAL AND AI-BASED CYBERSECURITY

Parameter	Traditional Security	AI-Based Security
Detection	Rules and signatures	Patterns and behaviour
Unknown threats	Often difficult	Potentially easier to identify
Data processing	Predefined indicators	Large-scale analysis
Adaptability	Rule updates required	Model retraining/updates
Analyst workload	High	Can reduce routine work
Explainability	Usually straightforward	May be difficult



Attack surface	Conventional controls	Includes model/data risks
Maintenance	Signatures and rules	Models, data and monitoring
Best use	Known threats and baseline controls	Anomaly analysis and prioritisation

VI. TRADITIONAL VS. AI-BASED CYBERSECURITY

Traditional and AI-based approaches should not be treated as mutually exclusive. Traditional mechanisms provide deterministic protection and remain highly effective for many known conditions. AI provides probabilistic analysis and can help discover patterns in large and changing datasets.

For example, a firewall can block traffic according to explicit policy, while an AI system can analyse whether a sequence of otherwise permitted connections is unusual. An antivirus signature can detect known malicious code, while behavioural analysis can identify suspicious execution even when a file has not previously been classified. The major limitation of AI is that it can make mistakes in ways that are difficult to anticipate. The model may also become less effective when the environment changes. Consequently, the practical architecture should combine deterministic controls, AI-based analytics, and human review.

A. Benefits of AI-Based Security

The principal benefits include faster analysis, large-scale processing, automation of repetitive tasks, anomaly identification, alert prioritisation, and the ability to correlate information from multiple sources. AI can also help security teams deal with information overload by reducing the number of alerts requiring immediate manual attention.

B. Limitations

AI-based security can suffer from false positives, false negatives, biased datasets, model drift, adversarial manipulation, high infrastructure requirements, and limited explainability. A model that performs well in a laboratory dataset may perform differently in a production environment. Evaluation should therefore include realistic operational conditions and continuous monitoring.

VII. PROPOSED LAYERED AI DEFENCE FRAMEWORK

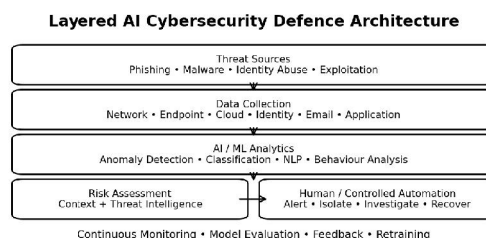


Fig. 2. Proposed layered AI cybersecurity defence architecture.

Based on the literature, a practical architecture can be organised into six layers: threat sources, data collection, AI/ML analytics, risk assessment, response, and continuous improvement. The design treats AI as an intelligence layer rather than as the only security control.

Fig. 1. Conceptual layered AI cybersecurity defence framework

CYBER THREATS → DATA COLLECTION → AI/ML ANALYSIS → RISK ASSESSMENT →
HUMAN/AUTOMATED RESPONSE → CONTINUOUS MONITORING & RETRAINING

The first layer represents threats such as phishing, malware, identity abuse, exploitation, and adversarial manipulation. The data layer collects network, endpoint, identity, application, cloud, email, and authentication telemetry. The analytics layer applies anomaly detection, classification, NLP, correlation, or other appropriate models. Risk assessment then combines model output with contextual information such as asset importance, user role, event history, and known threat intelligence. Response may include alerting, isolation, credential protection, blocking, investigation, or recovery. High-impact actions should normally require human approval.



The final layer continuously evaluates model performance and operational outcomes. Feedback can identify drift, false positives, new attack patterns, or changes in normal behaviour. This lifecycle is consistent with the broader risk-management approach promoted by NIST [1].

A. Human-in-the-Loop Principle

Human oversight is particularly important for security actions that can interrupt business operations or affect individuals. AI should present evidence, confidence, relevant context, and recommended actions so that analysts can make informed decisions. This design also provides an opportunity to detect systematic model errors.

B. Zero Trust Integration

AI can complement Zero Trust principles by analysing identity, device posture, application behaviour, and access context. However, AI should not replace explicit policy enforcement. Authentication, authorisation, least privilege, segmentation, and continuous verification remain core controls. NIST SP 800-207 describes Zero Trust Architecture as a model in which trust is not automatically granted based on network location [20].

VIII. REAL-WORLD THREAT CONTEXT AND CASE ANALYSIS

Recent threat reports show that AI is developing within an environment dominated by established attack methods. ENISA's 2025 Threat Landscape analysed approximately 4,875 incidents occurring from July 2024 through June 2025. It reported DDoS as the dominant incident type by incident count and identified ransomware as a particularly impactful threat [4]. This illustrates that organisations must continue addressing basic resilience and availability risks while adopting AI.

Phishing remains important because attackers often target identities rather than only technical vulnerabilities. AI-generated content can make social engineering more convincing, but strong identity controls, multi-factor authentication, conditional access, email protection, and user verification remain effective layers.

Ransomware demonstrates another lesson. Even if AI improves detection, organisations still require segmentation, backups, recovery testing, privileged-access controls, endpoint monitoring, and incident-response plans. AI can help identify suspicious behaviour, but resilience depends on the whole architecture.

Microsoft's 2025 Digital Defense Report describes phishing, identity attacks, ransomware, exposed services, and unpatched systems as important elements of the threat landscape [5]. This supports the conclusion that AI should supplement basic security hygiene rather than distract organisations from it.

IX. CHALLENGES AND LIMITATIONS

A. Data Quality and Availability

Security models require representative data. If training data contains bias, missing classes, outdated behaviours, or unrealistic laboratory conditions, performance may not transfer to production. Obtaining high-quality security datasets can also be difficult because organisations may not share incident data for privacy or commercial reasons.

B. False Positives and False Negatives

False positives increase analyst workload, while false negatives can allow attacks to remain undetected. The appropriate balance depends on the application. For a high-impact authentication decision, a false positive may have a different cost than for an automated malware quarantine. Evaluation should therefore consider operational consequences rather than accuracy alone.

C. Concept Drift

Normal behaviour and attack techniques change over time. A model trained on historical data may gradually lose accuracy. Continuous monitoring, periodic retraining, drift detection, and controlled model updates are required.

D. Adversarial Manipulation

Attackers may deliberately manipulate inputs, training data, or the model environment. Security systems should therefore be tested against adversarial conditions. NIST's adversarial ML taxonomy provides a common terminology for analysing these risks [2].



E. Explainability

Security analysts may need to understand why an event was classified as malicious. Explainability can improve trust and facilitate investigation, although explanations themselves must be evaluated for accuracy and usefulness. Complex models can create a trade-off between predictive capability and interpretability.

F. Privacy

AI security systems may process email content, authentication information, user behaviour, network metadata, and other sensitive information. Organisations should apply data minimisation, access controls, retention limits, encryption, and appropriate governance. Privacy should be considered during system design rather than after deployment.

G. Cost and Infrastructure

Large models and continuous analytics can require substantial computing resources. Organisations must consider infrastructure cost, latency, energy consumption, model maintenance, and the skills required to operate the system securely.

X. ETHICAL AND PRIVACY CONSIDERATIONS

AI in cybersecurity creates a tension between security monitoring and individual privacy. Monitoring can be justified for legitimate security purposes, but unnecessary collection can create new risks. Organisations should define what data is collected, why it is needed, who can access it, how long it is retained, and how it is protected.

Automated decisions also require accountability. If an AI system incorrectly blocks a legitimate user or labels an employee as suspicious, there should be a process for review and correction. NIST's AI RMF emphasises security, resilience, accountability, transparency, explainability, privacy, and fairness as important characteristics of trustworthy AI [1].

AI systems should also be protected against misuse by authorised users. Logging, role-based access control, secure APIs, secrets management, model governance, and audit procedures are therefore part of AI security itself.

A. Responsible Use

Responsible use includes clearly defined objectives, documented limitations, human oversight, security testing, privacy protection, and continuous monitoring. Organisations should avoid treating AI output as unquestionable truth. AI recommendations should be considered evidence that contributes to a decision, not necessarily the final decision.

B. Security of AI Supply Chains

AI applications may depend on third-party models, datasets, libraries, APIs, cloud services, and plugins. A weakness in any component can affect the overall system. Supply-chain security should therefore include dependency management, provenance, access control, update verification, and continuous monitoring.

XI. RESEARCH GAP

Several gaps remain in current research. First, many studies focus on using AI to detect attacks, while fewer operational studies examine the complete security lifecycle of AI models in production. Second, real-time adaptation remains difficult because models must respond to new behaviours without becoming unstable or learning from poisoned data. Third, explainable AI remains an open problem for complex security analytics.

Fourth, generative AI introduces new attack surfaces involving prompt manipulation, sensitive-data exposure, synthetic identities, model misuse, and autonomous agents. Fifth, there is a need for better methods to determine which actions can safely be automated and which require human approval. Finally, security research would benefit from more consistent benchmarks that represent real-world network conditions, evolving attacks, and adversarial behaviour.

A. Need for Standardised Evaluation

Comparing AI security models is difficult when studies use different datasets, attack definitions, evaluation metrics, and experimental environments. Future research should report common metrics such as precision, recall, F1 score, false-positive rate, false-negative rate, detection latency, resource consumption, and robustness under adversarial conditions. Where possible, evaluation should also include realistic deployment constraints.



XII. FUTURE SCOPE

Future research is likely to expand in several directions. Explainable AI can help analysts understand why a security model produced a particular result. Privacy-preserving ML can allow organisations to learn from distributed data without unnecessarily centralising sensitive information. Federated learning may be useful where multiple organisations need collaborative detection while limiting direct data sharing.

Adversarially robust ML is another important direction. Models should be tested not only for normal accuracy but also for resistance to manipulated inputs and compromised training data. Generative AI security will require research into prompt injection, data leakage, model abuse, secure agents, and trustworthy tool use.

AI-assisted Security Operations Centres may become more capable of summarising incidents and recommending response steps. However, controlled automation and human approval will remain important. AI can also support Zero Trust by analysing identity and device context, but policy enforcement should remain explicit and auditable.

Other promising areas include AI-assisted vulnerability management, synthetic-media detection, cloud security analytics, IoT anomaly detection, and continuous AI supply-chain monitoring.

A. AI and Cybersecurity Education

The future workforce will require both cybersecurity and AI skills. Security professionals need to understand model limitations, data pipelines, adversarial attacks, prompt-based risks, and AI governance. Conversely, AI engineers working on security-sensitive systems need knowledge of authentication, access control, threat modelling, secure development, and incident response. Interdisciplinary education is therefore an important research and professional-development area.

XIII. RECOMMENDATIONS

Based on the reviewed literature, organisations adopting AI for cybersecurity should follow these recommendations:

- 1) Use AI as one layer in a defence-in-depth architecture rather than as a replacement for conventional controls.
- 2) Secure the complete AI lifecycle, including datasets, training environments, models, APIs, deployment infrastructure, and logs.
- 3) Continuously monitor model performance and detect concept drift.
- 4) Perform adversarial testing before and after deployment.
- 5) Maintain human oversight for high-impact actions.
- 6) Use strong identity controls and multi-factor authentication because many attacks continue to target credentials.
- 7) Combine AI with patch management, network segmentation, secure configuration, backups, endpoint security, and vulnerability management.
- 8) Apply privacy-by-design principles and minimise unnecessary data collection.
- 9) Maintain audit trails so that AI-supported security decisions can be reviewed.
- 10) Evaluate systems using operationally meaningful metrics rather than relying only on accuracy.

A. Recommended Security Lifecycle

A practical lifecycle is: Identify assets and risks → Collect relevant telemetry → Analyse with AI and conventional controls → Validate findings → Respond according to risk → Review outcomes → Update models and controls. This creates a feedback loop in which lessons from incidents improve both human processes and technical controls.



XIV. DISCUSSION

AI as a Dual-Use Cybersecurity Technology



Fig. 3. AI as a dual-use technology in cybersecurity.

The reviewed evidence indicates that AI is changing cybersecurity in two directions simultaneously. Defensively, AI can process large quantities of information, identify patterns, prioritise alerts, and assist investigation. Offensively, AI can increase the speed and realism of some social-engineering and information-processing activities. The resulting environment is therefore not simply “AI versus attackers.” It includes conventional security controls, AI-enabled attacks, AI-enabled defence, human decision-making, and attacks against AI systems.

An important conclusion is that security teams should evaluate AI using the same risk-management principles applied to other critical technologies. AI models should have defined purposes, measurable performance requirements, secure deployment procedures, monitoring, access controls, and incident-response plans.

The value of AI is greatest when it addresses a clearly defined operational problem. For example, an organisation may gain significant benefit from using AI to reduce duplicate alerts or identify unusual authentication behaviour. By contrast, fully autonomous high-impact actions may create unacceptable operational risk without appropriate safeguards.

The literature also suggests that AI adoption should be gradual and measurable. Organisations can begin with analyst-assistance use cases, establish performance baselines, evaluate false-positive rates, and then introduce controlled automation. This approach supports human agency while still capturing the efficiency benefits of AI.

NIST's AI RMF and adversarial ML taxonomy reinforce the idea that AI risk must be managed across the lifecycle [1], [2]. The same principle applies to cybersecurity: model development, deployment, monitoring, and retirement all have security implications.

A. Overall Finding

The overall finding of this literature review is that AI is neither a universal solution nor merely an additional threat. It is a general-purpose technology that changes the capabilities available to both attackers and defenders. The most practical strategy is therefore a layered one that combines AI analytics with established security engineering and human expertise.

XV. CONCLUSION

Artificial Intelligence has become an important component of the modern cybersecurity ecosystem. The literature reviewed in this study demonstrates that AI and ML can assist with anomaly detection, intrusion detection, malware analysis, threat intelligence, behavioural analysis, phishing detection, fraud monitoring, and incident response. At the same time, AI can be misused to improve phishing, social engineering, synthetic identity attacks, vulnerability discovery, malware-related activities, and other malicious operations.

AI-based cybersecurity has significant limitations. Model effectiveness depends on data quality, security models can produce false positives and false negatives, models can become outdated, and adversarial techniques can manipulate AI behaviour. Privacy, explainability, accountability, governance, and human oversight are also important considerations. The research therefore supports a layered cybersecurity model. Organisations should combine AI-based detection and analysis with authentication, encryption, access control, patch management, network segmentation, backups, endpoint



protection, secure software development, and security awareness. AI should strengthen these controls rather than replace them.

Future cybersecurity will involve increasing collaboration between humans, AI systems, and conventional security technologies. The effectiveness of this collaboration will depend not only on algorithmic sophistication but also on secure engineering, representative data, continuous evaluation, responsible governance, and appropriate human oversight.

The central conclusion is that organisations should secure both ****systems with AI**** and ****the AI systems themselves****. Treating AI as both a security capability and a security-sensitive asset provides a balanced foundation for addressing the evolving cyber threat landscape.

TABLE II: AI-ENABLED ATTACKS AND DEFENSIVE RESPONSES

Attack Area	AI-Enabled Risk	Defensive Measures
Phishing	Personalised convincing content	Email security, MFA, awareness, behaviour analysis
Deepfake	Voice/video impersonation	Independent verification, MFA, contextual checks
Malware	Variant generation/evasion	Behavioural detection, sandboxing, endpoint telemetry
Adversarial ML	Manipulated inputs or data	Robust training, testing, monitoring
Vulnerability discovery	Faster weakness identification	Patching, secure SDLC, continuous assessment
Data poisoning	Corrupted training data	Data validation, provenance, access control
Credential abuse	Scaled social engineering	MFA, password managers, identity monitoring

TABLE III: AI SECURITY APPLICATIONS AND KEY LIMITATIONS

Application	Primary Benefit	Key Limitation
Anomaly detection	Identifies unusual behaviour	Legitimate anomalies cause alerts
IDS	Large-scale traffic analysis	Dataset drift and false positives
Malware detection	Detects known and unknown patterns	Adversarial evasion
Threat intelligence	Correlates large information volumes	Incorrect correlations
SOC assistance	Reduces analyst workload	Automation risk
Behavioural authentication	Context-aware access	Privacy and false positives
Fraud detection	Identifies unusual transactions	Changing user behaviour

REFERENCES

- [1] E. Tabassi, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD, USA, Jan. 2023, doi: 10.6028/NIST.AI.100-1.
- [2] A. Vassilev, A. Oprea, A. Fordyce, H. Anderson, X. Davies, and M. Hamin, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, NIST AI 100-2e2025, National Institute of Standards and Technology, Mar. 2025, doi: 10.6028/NIST.AI.100-2e2025.
- [3] M. Ozkan-Okay, E. Akin, Ö. Aslan, S. Kosunalp, T. B. Iliev, I. S. Stoyanov, and I. H. Beloev, “A comprehensive survey: Evaluating the efficiency of artificial intelligence and machine learning techniques on cyber security solutions,” IEEE Access, vol. 12, pp. 12229–12256, 2024, doi: 10.1109/ACCESS.2024.3355547.
- [4] European Union Agency for Cybersecurity (ENISA), ENISA Threat Landscape 2025, European Union Agency for Cybersecurity, Oct. 2025.
- [5] Microsoft, Microsoft Digital Defense Report 2025, Microsoft Security, 2025.
- [6] “A comprehensive review of AI’s current impact and future prospects in cybersecurity,” IEEE Access, vol. 13, pp. 14029–14050, 2025, doi: 10.1109/ACCESS.2025.3528114.
- [7] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Hoboken, NJ, USA: Pearson, 2021.
- [8] OWASP Foundation, OWASP Top 10: The Ten Most Critical Web Application Security Risks, 2021.



- [9] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1, 2024.
- [10] National Institute of Standards and Technology, Cybersecurity Framework (CSF) 2.0, NIST CSWP 29, 2024.
- [11] European Union Agency for Cybersecurity (ENISA), ENISA Threat Landscape for Supply Chain Attacks, 2021.
- [12] European Union Agency for Cybersecurity (ENISA), ENISA Threat Landscape 2024, 2024.
- [13] National Institute of Standards and Technology, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, NIST AI 100-2, 2023/2024.
- [14] IBM Security, Cost of a Data Breach Report 2025, IBM, 2025.
- [15] Cisco, 2025 Cybersecurity Readiness Index, Cisco, 2025.
- [16] Microsoft, Microsoft Digital Defense Report 2024, Microsoft Security, 2024.
- [17] IBM Security, X-Force Threat Intelligence Index 2025, IBM, 2025.
- [18] MITRE, MITRE ATT&CK: Enterprise Knowledge Base, MITRE Corporation.
- [19] National Institute of Standards and Technology, Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations, NIST SP 800-161 Rev. 1, 2022.
- [20] National Institute of Standards and Technology, Zero Trust Architecture, NIST SP 800-207, 2020.

