WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal Homepage: www.ijrpr.com ISSN 2582-7421

AI-Based Cybersecurity: A Risk-Aware Framework for Intelligent Threat Detection and Responsible Defense

Nagare Vaishnavi Pravin[1], Mr. R. R.Tajane[2] , Prof. Mahajan J.R.[3]

[1] Student, Department of Artificial Intelligence and Data Science, Adsul's Technical Campus, Chas, Ahilyanagar, Maharashtra, 414005, India
[2] Professor, Department of Artificial Intelligence and Data Science, Adsul's Technical Campus, Chas, Ahilyanagar, Maharashtra, 414005, India
[3] HOD, Department of Artificial Intelligence and Data Science, Adsul's Technical Campus, Chas, Ahilyanagar, Maharashtra, 414005, India

ABSTRACT

The growing dependence on cloud platforms, connected devices and digital services has increased the need for cybersecurity systems that can identify threats quickly and adapt to changing attack patterns. Artificial intelligence (AI) can support this goal through machine learning, deep learning, natural language processing and anomaly detection. However, AI-enabled defense also introduces risks, including biased training data, false alarms, adversarial manipulation, limited explainability and exposure of sensitive information through unapproved AI services. This paper reviews these opportunities and risks and proposes a risk-aware framework that combines multiple detection signals with human verification, privacy controls and continuous model monitoring. The framework is organized around data collection, feature analysis, risk scoring, analyst review, controlled response and feedback. It also considers "Shadow AI"—the use of AI tools outside an organization's approved governance—as a security issue that should be monitored alongside conventional network and endpoint threats. This is a literature-based conceptual study, not an experimental evaluation; therefore, no measured accuracy or performance improvement is claimed. The proposed framework offers a practical structure for organizations planning responsible AI-supported cybersecurity and identifies evaluation measures needed for future implementation.

Keywords: Artificial Intelligence, Cybersecurity, Machine Learning, Anomaly Detection, Phishing Detection, Explainable AI, Shadow AI, Human-in-the-Loop Security.

1. INTRODUCTION

Digital systems now support communication, finance, education, healthcare and public services. This dependence creates valuable targets for phishing, malware, account compromise, data theft and disruption. Conventional security controls—such as signatures, fixed rules and access policies—remain important, but they may not identify every unfamiliar or rapidly changing pattern. AI methods can assist analysts by processing large volumes of events, recognizing unusual behavior and prioritizing alerts [1, 2].

AI should not be treated as a replacement for established security controls or professional judgment. Its outputs depend on the data used for training, the environment in which it is deployed and the way alerts are acted upon. A model may miss a new attack, flag legitimate activity as suspicious or be manipulated by carefully crafted inputs. In addition, staff may use unapproved AI applications to process organizational information, creating data leakage and compliance concerns [3].

This paper develops a risk-aware, conceptual framework for AI-supported cybersecurity. Its focus is the combination of threat detection with governance and response safeguards. The main contributions are: (i) a structured overview of relevant AI techniques and their roles; (ii) a workflow that combines technical signals with risk and confidence checks; (iii) safeguards for explainability, privacy and human oversight; and (iv) an evaluation plan that can be used in a future prototype without inventing experimental results.

2. PROBLEM STATEMENT AND OBJECTIVES

Many organizations need to identify suspicious activity across network traffic, user behavior, endpoints, messages and cloud services. These sources differ in format and reliability, while security teams may face large alert volumes and limited investigation time. A single model or isolated tool may not provide enough context to distinguish a real attack from an unusual but legitimate event. Meanwhile, unauthorized AI tools can create a separate pathway for sensitive data to leave controlled environments [2, 3].

The objective is to outline a modular AI-based cybersecurity framework that supports detection and prioritization while retaining human accountability. The specific objectives are to: (1) map AI techniques to common security tasks; (2) combine model confidence with contextual risk indicators; (3) include

privacy and governance controls for AI tool usage; (4) describe safe response stages; and (5) define measurable criteria for evaluating a future implementation.

3. LITERATURE REVIEW

Al Siam, Hassan and Bhuiyan survey AI methods used in cybersecurity, including machine learning for behavioral analysis, deep learning for complex patterns, natural language processing for language-based threats and anomaly detection for unusual activity [1]. Their discussion emphasizes that methods have different strengths and constraints: deep models can require substantial computing resources, language models may struggle with changing social-engineering tactics, and anomaly detectors can produce false positives.

The paper "Exploring the Impact of Artificial Intelligence on Cyber Security" discusses the dual role of AI: it can improve threat detection, vulnerability management and incident response, but also raises concerns about data quality, adversarial attacks, explainability, scalability and regulation [2]. These issues indicate that deployment quality depends on more than a model's predictive score; operational safeguards and analyst workflows matter as well.

Puthal and colleagues examine Shadow AI, describing how unapproved AI services may process sensitive data outside formal IT oversight. Their review identifies risks such as data leakage, excessive permissions, insecure third-party integrations and weaknesses in governance. It also discusses possible defensive uses of AI monitoring, explainable AI and privacy-aware approaches [3]. This broadens the security problem from detecting external attackers to also managing the organization's own use of AI tools.

Across these sources, a recurring gap is the need to consider detection, response and governance as connected parts of one workflow. This paper addresses that design gap conceptually by placing risk scoring, human review and controlled response around AI-generated alerts. The proposal is a synthesis of the supplied literature, not a claim that the framework has already been implemented or experimentally validated.

4. REVIEW METHODOLOGY

This study uses a focused narrative review of the three supplied research papers: a 2025 conference paper on AI methods for cybersecurity [1], a review of AI's cybersecurity opportunities and challenges [2], and a 2025 survey on Shadow AI risks and opportunities [3]. The documents were examined for recurring AI techniques, security use cases, limitations and governance concerns. Findings were grouped into four themes: detection capabilities, operational limitations, AI-specific risks and safeguards. No new dataset was collected, no model was trained and no quantitative experiment was conducted. The resulting framework should therefore be understood as a conceptual design for future testing.

5. PROPOSED RISK-AWARE FRAMEWORK

The framework contains six connected stages: (1) data collection, (2) feature and context analysis, (3) multi-signal risk assessment, (4) analyst verification, (5) controlled response, and (6) feedback and governance. It is designed to complement existing firewalls, endpoint protection, identity controls and security information and event management systems rather than replace them.

5.1 Data Collection and Preparation

The system may receive security logs, authentication events, endpoint alerts, network-flow summaries, email metadata and approved indicators of compromise. Data collection should follow least-privilege access, retention limits and applicable privacy requirements. Before analysis, records should be normalized, timestamps aligned, duplicate events reduced and sensitive fields masked where possible. Data provenance and labeling quality should be recorded so that model decisions can be audited.

5.2 AI-Based Detection Layer

Different techniques can be assigned to tasks suited to their input. Supervised machine-learning models can classify events when representative labeled examples are available. Unsupervised or semi-supervised anomaly detection can flag deviations where labels are scarce. Deep-learning methods may analyze complex event sequences or malware features, while natural language processing can help assess suspicious messages and phishing-related text [1, 2]. The system should not assume that any one technique detects all threats; model outputs should be treated as evidence rather than certainty.

5.3 Contextual Risk Scoring

Each alert can be prioritized using several factors: model confidence, asset criticality, user privilege, novelty of the behavior, severity of the suspected activity and corroborating signals from independent controls. A simple conceptual score can be written as $R = w_1S + w_2C + w_3A + w_4N$, where S represents severity, C confidence, A asset or account criticality, N novelty, and w_1–w_4 are weights selected and validated for the organization. This expression describes a design option, not a tested scoring model. Weights should be calibrated using historical incidents and reviewed for bias and drift.

5.4 Human Verification and Controlled Response

High-impact actions—such as disabling an account, isolating a business-critical host or blocking a major service—should require an authorized analyst or a predefined approval policy. Lower-risk actions, such as enriching an alert or requesting additional logs, may be automated. Every action should preserve an audit trail containing the alert source, supporting evidence, confidence, decision, approver and outcome. This human-in-the-loop approach reduces the chance that a mistaken model prediction directly causes a disruptive response.

5.5 Shadow AI Governance

Organizations should maintain an inventory of approved AI services, define what information may be submitted to them and monitor for unauthorized integrations. Access controls, data-loss prevention, vendor review, retention rules and employee guidance can reduce accidental disclosure. Detection of possible Shadow AI use should be handled proportionately: metadata and policy signals may justify review, but monitoring should respect privacy and employment rules. Where an AI service is useful, a secure approval route is preferable to relying only on prohibition [3].

5.6 Feedback and Model Maintenance

Analyst-confirmed outcomes can support periodic retraining and rule refinement, subject to data quality and privacy checks. Model updates should be tested against held-out data and representative operational conditions before deployment. Monitoring should track performance drift, false-positive rates, missed incidents, response time and changes in the data distribution. Rollback procedures should be available if an update degrades security outcomes.

6. DISCUSSION

The main value of the proposed design is the connection between detection and operational decision-making. Machine learning, deep learning, NLP and anomaly detection provide complementary signals, while risk scoring helps order alerts for investigation [1]. Human review and audit records make the resulting decisions more accountable. Governance for Shadow AI is included because a secure organization must protect against attacks and manage the risks created by its own unapproved technology use [3].

The framework also has important limitations. Poor or unrepresentative training data can lead to biased predictions. A changing environment can increase false positives or reduce detection of new attacks. Attackers may attempt to evade or manipulate models, and explainability techniques do not guarantee that a prediction is correct. Integrating multiple tools can increase cost and complexity, particularly for smaller organizations. Consequently, AI outputs should be combined with established security controls, tested against realistic scenarios and reviewed by qualified personnel [1, 2].

7. EVALUATION PLAN

A future implementation should be evaluated using a documented dataset, a clearly separated test set and a repeatable test environment. The dataset should describe its source, time range, class balance, labeling procedure and any privacy-preserving transformations. Where feasible, the test split should be separated by time or source to reduce leakage between training and evaluation data.

Recommended measures include precision, recall, F1-score and false-positive rate for detection; time-to-detect and time-to-respond for operations; and resource consumption for deployment. Results should be reported per threat category as well as overall, because aggregate scores can hide weak performance on rare but important events. Robustness tests should examine changed traffic patterns, previously unseen examples and controlled adversarial perturbations. A separate governance assessment should measure policy violations, unauthorized AI integrations, audit completeness and the proportion of high-impact actions receiving required approval.

No numerical results are reported in this paper because no implementation or experiment was performed. Any future paper version should replace this evaluation plan with measured values, dataset details, comparison baselines and limitations observed during testing. Illustrative or expected values must not be presented as experimental findings.

8. LIMITATIONS AND FUTURE SCOPE

This work is based on a focused review of three supplied papers and is not an exhaustive systematic review of all AI cybersecurity research. The proposed architecture is conceptual and its risk score, thresholds and governance procedures have not been validated experimentally. The reviewed sources may also cover different environments and threat assumptions, so their findings should not be treated as directly comparable benchmark results.

Future work can implement a prototype that integrates log analysis, phishing-text classification and anomaly detection into a single alert workflow. Research should compare individual models with combined approaches, test explainability methods with analysts, and evaluate privacy-preserving collaboration such as federated learning where appropriate. Further work should also investigate model poisoning, prompt-based attacks against AI-enabled security assistants and methods for identifying unauthorized AI services without excessive employee surveillance.

9. CONCLUSION

AI can strengthen cybersecurity by supporting behavioral analysis, anomaly detection, phishing identification and prioritization of security events. However, its effectiveness depends on data quality, robustness, interpretability and the way predictions are used in operational decisions. This paper has proposed a risk-aware framework that combines multiple detection techniques with contextual scoring, human verification, controlled response, feedback and Shadow AI governance. The framework is a conceptual synthesis of the supplied literature, not a validated product or measured experiment. Its next step is implementation and evaluation using representative data, transparent metrics and documented safety controls.

REFERENCES

[1] Al Siam, A., Hassan, M. M., & Bhuiyan, T. (2025). Artificial Intelligence for Cybersecurity: A State of the Art. Proceedings of the 4th IEEE International Conference on AI in Cybersecurity (ICAIC), University of Houston, Texas, USA. https://doi.org/10.1109/ICAIC63015.2025.10848980

[2] Devansh, Arora, D., Tyagi, P., & Dadhich, P. (n.d.). Exploring the Impact of Artificial Intelligence on Cyber Security: Challenges, Opportunities, and Future Trends. SSRN Electronic Journal. https://ssrn.com/abstract=4837827

[3] Puthal, D., Mishra, A. K., Mohanty, S. P., Longo, A., & Yeun, C. Y. (2025). Shadow AI: Cyber Security Implications, Opportunities and Challenges in the Unseen Frontier. SN Computer Science, 6, 405. https://doi.org/10.1007/s42979-025-03962-x