Journal of Media Horizons
ISSN: 2710-4060 2710-4052 Volume 7, Issue 6, 2026

AI-DRIVEN CYBERSECURITY: THREAT DETECTION AND MITIGATION STRATEGIES

Faizan Sagheer[*1], Muhammad Mubashir Gujjar[2], Muhammad Usman Akhtar[3]

[*1]*Department of Computer Science, Islamia University of Bahawalpur, Bahawalpur, Punjab, Pakistan*
[2]*Department of information Sciences, University of Education, Lahore, Dera Ghazi Khan, Punjab, Pakistan*
[3]*Cadet College Hasanabadal, Hasanabdal, Punjab, Pakistan*

[*1]faizansagheer09@gmail.com

DOI: https://doi.org/10.5281/zenodo.20699713

Keywords
Artificial intelligence; cybersecurity; threat detection; mitigation strategies; machine learning; cyber threats; anomaly detection; incident response; ROC-AUC; network security

Article History
Received: 18 April 2026
Accepted: 30 May 2026
Published: 15 June 2026

Corresponding Author: *
Faizan Sagheer

Abstract

The increasing frequency, complexity, and operational impact of cybersecurity threats necessitate intelligent systems that can quickly detect and effectively remediate threats. Using an alert-level analytical design, this study assessed an AI-enabled cybersecurity framework for recognizing and mitigating cyber threat events. 500 Cybersecurity alert records were analyzed; both threat and non-threat events. The diagnostic performance of the AI-based model was evaluated using accuracy, sensitivity, specificity, positive and negative predictive values, and receiver operating characteristic curve analysis. Chi-square tests were used to assess correlations between cybersecurity indicators and cyber-threat status, and logistic regression was used to evaluate predictors of cyber-threat occurrence. Of the 500 analyzed alerts, 179 were actual cyber-threat events, or a 35.8% overall threat incidence. This AI-based model achieved an overall accuracy of 91.0%, sensitivity of 74.9%, specificity of 100.0%, positive predictive value of 100.0%, and negative predictive value of 87.7%. Then, the ROC-AUC was 0.997, indicating perfect discrimination. Actual threat status was most significantly associated with patch status, strange port access, geo-anomaly, and signature match. It indicates that threat detection, alert prioritization, and timely mitigation with autonomous decision-making capability can be improved by AI-driven systems. This result clearly demonstrates the need for continuous optimization of classification, threshold adjustment, and expert review to create stronger overall detection and improve threat capture.

1. INTRODUCTION

The rapid growth of digital technologies, cloud computing, Internet of Things systems, digital financial platforms, and interlocked organizational networks has made cybersecurity threats more complex and frequent (Ayodele et al., 2024). Cyberattacks have evolved from just unauthorized access to information to also include malware, phishing, distributed denial-of-service (DDoS) attacks, brute-force intrusions, insider anomalies, ransomware-related acts, and advanced persistent threats. These attacks can result in compromised data confidentiality, availability degradation, reputation loss as well as financial and operational losses (Obi et al., 2024). Despite the increasing scale and sophistication of cyber threats, legacy rule-based security systems encounter difficulties in detecting barely visible, advanced, or rapidly changing attack patterns (Ganesan et al., 2019).

Artificial intelligence is rapidly emerging as a major avenue towards strengthening detection and response to cybersecurity threats. By analyzing large volumes of network traffic, endpoint behavior, login patterns, access activity, and indicators of anomalies in real-time, AI systems can help organizations quickly detect possible intrusions after they occur. With traditional security solutions dependent on static signatures, AI-based models can help to identify anomalous patterns, detect anomalous behavior, classify alerts, and predict threats ahead of time before they can become threats (Shanthi et al., 2023). By filtering through massive numbers of alerts and redefining threat classification, machine learning and smart analytics can alleviate the pressure on Cybersecurity analysts. Nevertheless, recognizing the urgency of timely mitigation decisions, these solutions are only valuable when they achieve high accuracy and overall low rates in both false positives and false negatives (Ghadermazi et al., 2024).

Threat mitigation is very important in cybersecurity because detecting threats is not enough without response actions. Some remediation strategies include denial of access, endpoint isolation, credential resets, blocking, constant monitoring, and escalation to incident response teams (Leventopoulos et al., 2024). Factoring AI into cybersecurity operations can bolster mitigation efforts, as organizations can use AI to zero in on high-priority alerts, identify appropriate response strategies, reduce response time, and minimize operational disruption. Thus, this research sought to evaluate an AI-based cybersecurity framework for threat detection and mitigation in terms of detection performance, examine the leading security indicators most susceptible to cyber-threat events, and analyze the interactions between mitigation strategies and cybersecurity response measures (ABEL et al., 2024).

2. Methodology

2.1 Study Design

This was an analytical study conducted at Department of Computer Science, Islamia University of Bahawalpur, Bahawalpur, Punjab, Pakistan to see How an AI-based cybersecurity framework works for detecting or mitigating threats. They used an Alert-level analytical method in which each observation represented a cybersecurity event detected by an AI-based detection system.

2.2 Study Variables

The outcome variable assessed in this study was whether the threat status was real (threat) or not (non-threat). Threat events involved malware, phishing, distributed denial-of-service attacks, brute-force attempts, insider anomalies, and ransomware prep work. Predictor Variables: AI threat score, failed logins, endpoint risk score, patching status, geo-anomaly, unusual port access, signature match, traffic volume, user privilege level, and device type.

2.3 AI-Based Threat Detection

The AI-powered framework categorized cybersecurity alerts by their probability score. The alerts that scored higher fall under the suspicious or malicious category. The performance of the detection was compared with the actual threat status detected in both methods: AI-generated classification vs. real threat status. The classification results were categorized into true positive, true negative, false positive, and false negative.

2.4 Mitigation Strategy Assessment

Response time, containment status, downtime, and data exfiltration indicator were used to evaluate the mitigation effectiveness. Mitigation actions such as monitoring, access restrictions, endpoint isolation, credential resets, traffic blocking, and incident escalation were included. Successful mitigation was defined as containing the threat, but no evidence of exfiltrated data.

2.5 Statistical Analysis

We used IBM SPSS Statistics to analyze data. For all study variables, descriptive statistics were computed. Frequencies and percentages for categorical variables and means and standard deviations for continuous variables were calculated.

Chi-square tests were used to assess the association between categorical predictors and actual threat status. We conducted binary logistic regression to identify predictors of cyber-threat occurrence. Outcomes were expressed as odds ratios with 95% confidence intervals. Receiver operating characteristic curve analysis was performed for assessment of discrimination of AI threat scores. Model performance was assessed using the accuracy, sensitivity, specificity, positive and negateve predictive value, and area under curve. Statistical significance was defined as p-values < 0.05.

3. Results

3.1. Distribution of Cybersecurity Alerts

We examined 500 records of cybersecurity alerts. Out of these alerts, 179 alerts (35.8%) belong to cyber-threat event type, while 321 alerts (64.2%) were classified as non-threat events. Of these alerts, the AI-based system responded with 134 alerts as threat alerts and 366 alerts as non-threat alerts.

Variable	Category	Frequency	Percentage
Total alerts analyzed	—	500	100.0
Actual threat status	Threat	179	35.8
Actual threat status	Non-threat	321	64.2
AI detection status	Detected as threat	134	26.8
AI detection status	Detected as non-threat	366	73.2

3.2. Distribution of Cyber-Threat Types

Of the 179 alerts classified as true threats, the documented threat types consisted of malware, phishing, distributed denial-of-service attacks, brute-force attempts, insider anomalies and ransomware precursor activity.

3.3. AI Detection Outcomes

In the comparison of actual threat status along with AI classification, there were 134 true positive alerts, 321 true negative alerts, 45 false negatives, and no false-positive classification alerts were recorded. The results suggest that the AI model was very specific but not sensitive; many of the cyber-threat events it flagged as non-cyber-threat events were actually true positives.

Detection outcome	Frequency	Percentage
True positive	134	26.8
True negative	321	64.2
False negative	45	9.0
False positive	0	0.0
Total	500	100.0

3.4. Diagnostic Performance of the AI-Based Threat Detection Model

The AI-driven threat detection model achieved an overall accuracy of 91.0%. Sensitivity was 74.9%, indicating that the model correctly identified approximately three-fourths of actual cyber-threat events. Specificity was 100.0%, showing that all non-threat events were correctly classified as non-threats. The positive predictive value was 100.0%, while the negative predictive value was 87.7%.

Performance metric	Value
Accuracy	91.0%
Sensitivity	74.9%
Specificity	100.0%
Positive predictive value	100.0%
Negative predictive value	87.7%
False-negative rate	25.1%
False-positive rate	0.0%
ROC-AUC	0.997
95% CI for ROC-AUC	0.994–1.000

3.5. Receiver Operating Characteristic Curve Analysis

Receiver operating characteristic curve analysis showed excellent discriminatory performance of the AI threat score. The area under the curve was 0.997, with a 95% confidence interval of 0.994–1.000. This indicates that the AI threat score had a strong ability to distinguish threat alerts from non-threat alerts.

3.6. Association Between Cybersecurity Indicators and Actual Threat Status

The chi-square test showed that patch status, unusual port access, geo-anomaly, and signature match were significantly associated with actual cyber-threat status. Signature match showed the strongest association with threat classification, followed by unusual port access, patch status, and geo-anomaly.

Predictor variable	Statistical test	Test value	p-value	Interpretation
Patch status	Chi-square test	87.71	<0.001	Significant association
Unusual port access	Chi-square test	113.25	<0.001	Significant association
Geo-anomaly	Chi-square test	85.72	<0.001	Significant association
Signature match	Chi-square test	226.86	<0.001	Strong significant association

3.7. Predictors of Cyber-Threat Occurrence

The results of binary logistic regression analysis revealed that owning a higher endpoint risk score, higher failed login attempts, unusual port access, geo-anomaly and signature match was positively associated with the probability of correct cyber-threat classification. Threat occurrence was also associated with patch status.

Predictor variable	Direction of association	Interpretation
Endpoint risk score	Positive	Higher endpoint risk increased the likelihood of actual threat occurrence
Failed login attempts	Positive	Increased failed login attempts were associated with higher threat likelihood
Unusual port access	Positive	Presence of unusual port activity increased the likelihood of threat occurrence
Geo-anomaly	Positive	Geographic anomaly was associated with increased threat likelihood
Signature match	Positive	Signature match strongly increased the likelihood of actual threat classification
Patch status	Associated	Unpatched or outdated systems were more frequently linked with threat events

3.8. Mitigation Strategy Distribution and Response Assessment

Mitigation tactics involved monitoring, restricting access, isolating the endpoint, resetting user credentials, blocking traffic and escalating the incident. Response time, containment status, downtime, and indicators of data exfiltration were also evaluated as part of mitigation response.

Mitigation variable	Description	Outcome relevance
Mitigation strategy	Action applied after detection	Evaluated the type of response used
Response time	Time from alert generation to mitigation initiation	Measured response efficiency
Containment status	Whether the threat was controlled	Assessed mitigation success
Downtime	Operational disruption following the alert	Measured system impact
Data exfiltration indicator	Evidence of unauthorized data transfer	Assessed severity of security failure

4. Discussion

We explored an AI-based cybersecurity framework for threat detection and mitigation in this research. The performance was shown to be very high overall, with excellent specificity and a ROC-AUC indicating very good discrimination between threat and non-threat alerts. This suggests that the AI system performed very well in accurately marking nonthreat events with high specificity and positive predictive value, thus reducing excessive false-positive alarms. This is a very important in cybersecurity operations because too many false positives can present unnecessary burdens, slow down response times, and erode trust in the automated detection deployment (Almarzooqi et al., 2025).

The model performed well overall, but sensitivity was lower than specificity. This means that some cyber threats were not detected because false-negative cases were observed. False negatives are especially damaging, as a threat could result in system compromise, data exfiltration, service disruption, or network compromise without detection in real-world cybersecurity ecosystems (Erbacher, 2022, Muoio, 2023). Thus, while all indications are that the model can be trusted to report detected threats, it might need further fine-tuning in identifying all types of malicious activity especially when they are stealthy or in the early stages of infection, such as ransomware precursors, insider anomalies, and low-frequency intrusion attempts (Wang et al., 2024).

The association analysis identified patch status, unusual port access, geo-anomaly, and signature match as being significantly associated with the true cyber-threat status. These results substantiate the approach of including system-level vulnerabilities, behavioral deviations, and known threat signatures when creating an AI-based cybersecurity model. The highest correlation to actual threat status came from signature match (Ara, 2026, Chen et al., 2024). Additionally, while geo-anomaly and east-west port access rely on both anomaly-based detection and a baseline of pre-existing threat signatures, they equally underscore the importance of anomaly-based detection for recognizing suspicious activity. This justifies combining signature and behavior detection using a hybrid AI approach (Chen et al., 2024).

The mitigation findings provide evidence that AI-driven threat detection can support timely, risk-based cybersecurity incident response strategies. Alerts linked to additional threat indicators could result in more robust mitigation steps, including endpoint sanitization, access blocking, credential resets, traffic blocking, or escalation to incident response groups. Detecting an event accurately is not enough; however, rapid response, proper containment, and minimization of impact on operations are key to effective and efficient mitigation. In conclusion, the results suggest that AI-enabled cybersecurity systems can help

prioritize threats and inform mitigation decisions. At the same time, modeling and training still require continual updating, validation, threshold adjustments, and human expert review to minimize missed threats and improve real-world operational performance (Narayan et al., 2025, Thapaliya, 2025).

5. Conclusion

The results show that AI-based cybersecurity framework is effective in threat detection and protection through threat and non-threat alert classification and critical security indicator identification of cyber-threat occurrence. The model achieved overall high accuracy, good specificity and model robustness, suggesting its potential value to minimize false-positive alerts and enhance surveillance monitoring efficiency. Nevertheless, the occurrence of false negatives underscores the necessity for ongoing model improvement, threshold adjustment, and collaboration with human experts to mitigate missed threats. Strong associations of actual threat status with patch status, abnormal port utilization, geolocation-anomaly and signature match reinforce that a hybrid of vulnerability-based, signature-based and anomaly-based features is critical in cyber threat analytics. Overall, AI-based threat detection systems can enhance cyber defense strategies by improving alert prioritization, supporting timely mitigation decisions, and strengthening organizational resilience against evolving cyber threats.

6. Funding

Not applicable.

7. References

ABEL, U., EMMANUEL, C. & PASCAL, U. O. 2024. Applying artificial intelligence in Cybersecurity to enhance threat detection, response, and risk management. *COMPUTER SCIENCE,* 5**,** 2511-2538.

ALMARZOOQI, H., SHAALAN, K., ALHASHMI, M. & ALHARTHI, S. Artificial Intelligence in Cyber Threat Intelligence: A Systematic Review of Techniques and Applications. 2025 3rd International Conference on Cyber Resilience (ICCR), 2025. IEEE, 1-11.

ARA, A. 2026. Enhancing E-Commerce Security: A Framework of Integrated Vulnerability Assessment and Data Privacy Protection. *AI-Powered Content Delivery Networks***,** 213.

AYODELE, J., JAVED, A. & PORIETE, A. Understanding the impact of risk perception on cybersecurity training effectiveness in small and medium enterprises (SMEs). International Conference on Cyber Security, Privacy in Communication Networks, 2024. Springer, 103-112.

CHEN, H., SHEN, Z., WANG, Y. & XU, J. 2024. Threat detection driven by artificial intelligence: Enhancing cybersecurity with machine learning algorithms. *World J. Innov. Mod. Technol.,* 7**,** 58-70.

ERBACHER, R. F. 2022. Base-rate fallacy redux and a deep dive review in cybersecurity. *arXiv preprint arXiv:2203.08801*.

GANESAN, A., PARAMESHWARAPPA, P., PESHAVE, A., CHEN, Z. & OATES, T. 2019. Extending signature-based intrusion detection systems withbayesian abductive reasoning. *arXiv preprint arXiv:1903.12101*.

GHADERMAZI, J., SHAH, A. & JAJODIA, S. 2024. A machine learning and optimization framework for efficient alert management in a cybersecurity operations center. *Digital Threats: Research and Practice,* 5**,** 1-23.

LEVENTOPOULOS, S., PIPYROS, K. & GRITZALIS, D. 2024. Retaliating against cyber-attacks: a decision-taking framework for policy-makers and enforcers of international and cybersecurity law. *International Cybersecurity Law Review,* 5**,** 237-262.

MUOIO, P. 2023. Assessing Risk in Cybersecurity: How Sound Data Science Can Raise the Bar. *CHANCE,* 36**,** 4-8.

NARAYAN, Y., GUPTA, A., CHARAN, P., LAKSHMI, T. H., GHUMMAN, M. K. & SAGHRA, H. S. Strengthening Network Security through the Implementation of AI-Driven Automated Incident Response Systems. 2025 IEEE International Conference on Compute, Control, Network & Photonics (ICCCNP), 2025. IEEE, 1-5.

OBI, O. C., AKAGHA, O. V., DAWODU, S. O., ANYANWU, A. C., ONWUSINKWUE, S. & AHMAD, I. A. I. 2024. Comprehensive review on cybersecurity: modern threats and advanced defense strategies. *Computer Science & IT Research Journal,* 5**,** 293-310.

SHANTHI, R. R., SASI, N. K. & GOUTHAMAN, P. A new era of cybersecurity: the influence of artificial intelligence. 2023 international conference on networking and communications (ICNWC), 2023. IEEE, 1-4.

THAPALIYA, S. 2025. Artificial Intelligence and Cybersecurity: Pioneering Next-Generation Protection Strategies. *SADGAMAYA,* 2**,** 61-65.

WANG, Z., ZHOU, Y., LIU, H., QIU, J., FANG, B. & TIAN, Z. 2024. Threatinsight: Innovating early threat detection through threat-intelligence-driven analysis and attribution. *IEEE Transactions on Knowledge and Data Engineering,* 36**,** 9388-9402.