INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 4 • November 2026

EXPLAINABLE AGENTIC AI FOR REAL-TIME CYBERATTACK DETECTION AND AUTONOMOUS RESPONSE IN PAKISTAN'S CRITICAL DIGITAL INFRASTRUCTURE

Ali Hussnain[*1], Naseer Ahmad[2], Munwar Ali[3], Bhavnesh Deep[4], Asad Ullah[5]

[*1]*Student, Riphah International University*
[2]*Ph.D Data Science Scholar, Institute of Data Science, University of Engineering and Technology Lahore*
[3]*Associate Professor, Department of Computer Science, Shaheed Benazir Bhutto, University Shaheed Benazirabad*
[4]*Software Engineer, Graduate, Department of Computer Science, SZABIST University, Karachi*
[5]*Student, Institute of Computer Science and Information Technology, The University of Agriculture Peshawar, Khyber Pakhtunkhwa, Pakistan, 25130*

[*1]alihussnainpersonal@gmail.com, [2]naseer.ahmad@uet.edu.pk, [3]munwar.ali@sbbusba.edu.pk, [4]bhavneshkarmani485@gmail.com, [5]asadjn99@gmail.com

DOI: https://doi.org/10.5281/zenodo.24299848

Keywords
Agentic artificial intelligence, Explainable AI, Cyberattack detection, Autonomous cyber defence, Critical infrastructure, Human oversight.

Article History
Received: 08 August 2026
Accepted: 22 October 2026
Published: 17 November 2026

Copyright @Author
Corresponding Author: *
Ali Hussnain

Abstract

The increasing digitalization of Pakistan's critical infrastructure has intensified exposure to sophisticated cyber threats, highlighting the limitations of conventional intrusion detection and response systems. This study proposed an explainable agentic artificial intelligence (AI) framework for real-time cyberattack detection and autonomous response. The framework integrated machine-learning-based threat detection, explainable AI techniques (SHAP and LIME), coordinated intelligent agents, and human-supervised response mechanisms. A quantitative experimental design was proposed using publicly available cybersecurity datasets, including CICIDS2017, UNSW-NB15, and ToN-IoT, within a controlled cyber-range environment. Detection performance was evaluated through accuracy, precision, recall, F1-score, and false-positive rate, while operational effectiveness was assessed through detection latency, response time, threat containment, and service continuity. The results indicated potential improvements in detection performance, response efficiency, and operational reliability compared with conventional intrusion detection approaches. The study contributes to socio-technical systems theory by emphasizing the integration of intelligent automation, explainability, and human oversight in cybersecurity. The proposed framework offers a potential foundation for strengthening cyber resilience across Pakistan's critical digital infrastructure.

Introduction

The increasing digitalization of banking, telecommunications, energy, healthcare, and government services has exposed critical infrastructure to sophisticated cyber threats. Conventional intrusion detection systems often face challenges in identifying previously unknown attacks, minimizing false alarms, and responding to threats in real time. Artificial intelligence (AI) offers advanced capabilities for detecting abnormal network behaviour, while explainable AI (XAI) improves the transparency and interpretability of automated security decisions (Ali & Sagiroglu, 2025; Mohale & Obagbuwa, 2025).

Recent developments in agentic AI have introduced the potential for autonomous cyber defence through intelligent agents capable of

monitoring networks, analysing threats, coordinating decisions, and executing authorized response actions. Loevenich et al. (2025) demonstrated a hybrid autonomous cyber-defence architecture combining deep reinforcement learning, large language models, and rule-based systems. Similarly, Morris et al. (2025) highlighted the potential of reinforcement-learning agents for timely and autonomous cyber defence. However, reliable explanations, operational safety, and human oversight remain essential for deploying autonomous systems in critical environments.

In Pakistan, the National Cyber Security Policy 2021 emphasizes the protection of critical information infrastructure, institutional cybersecurity capabilities, and coordinated incident response (Ministry of Information Technology and Telecommunication [MoITT], 2021). The Computer Emergency Response Teams Rules 2023 further establish mechanisms for national, sectoral, and organizational cybersecurity coordination (Government of Pakistan, 2023).

Against this background, this study proposes an integrated explainable agentic AI framework for real-time cyberattack detection and controlled autonomous response in Pakistan's critical digital infrastructure. It focuses on detection performance, decision explainability, response efficiency, and human-supervised operational safety.

Problem Statement

Pakistan's increasing dependence on interconnected digital infrastructure has intensified the need for effective cyberattack detection and rapid incident response. Critical sectors, including banking, telecommunications, energy, and government services, require reliable cybersecurity mechanisms to protect sensitive information and maintain uninterrupted operations.

Although machine-learning-based intrusion detection systems have improved threat identification, challenges remain in detecting novel attacks, reducing false positives, explaining automated decisions, and responding effectively to evolving threats. Furthermore, conventional security systems frequently depend on manual investigation and intervention, potentially delaying incident containment. Agentic AI offers opportunities to automate threat analysis and response; however, insufficient explainability, unreliable autonomous decisions, and the possibility of disrupting essential services create significant deployment concerns (Loevenich et al., 2025; Mohale & Obagbuwa, 2025).

The principal research gap concerns the limited integration of explainable threat detection, coordinated agentic decision-making, and controlled autonomous response within a unified cybersecurity framework. In Pakistan, context-specific technical evidence concerning the effectiveness, reliability, and operational safety of such integrated systems remains limited.

Therefore, this study investigates how explainable agentic AI can improve real-time cyberattack detection and autonomous response while maintaining human oversight and operational accountability. It aims to develop and experimentally evaluate a framework suitable for the security requirements of Pakistan's critical digital infrastructure.

Research Questions

- RQ1: How effectively can explainable agentic AI detect cyberattacks in real time?
- RQ2: To what extent does explainability improve the transparency and reliability of AI-based threat detection?
- RQ3: How does autonomous AI response affect incident containment time and response effectiveness?
- RQ4: How does human oversight influence the safety and reliability of autonomous cyber defence?
- RQ5: How effectively can the proposed framework support the cybersecurity requirements of Pakistan's critical digital infrastructure?

Research Objectives

- To develop an explainable agentic AI framework for real-time cyberattack detection.
- To evaluate its threat-detection performance using accuracy, precision, recall, F1-score, and false-positive rate.

- To assess the contribution of explainability to the transparency and reliability of threat classification.
- To measure the effectiveness of autonomous response in reducing incident containment time.
- To evaluate the role of human oversight in ensuring safe and accountable autonomous responses.
- To assess the framework's applicability to Pakistan's critical digital infrastructure.

Significance of the Study

Theoretical significance: Integrates explainable AI, agentic decision-making, and human oversight into an integrated framework for autonomous cybersecurity, extending the application of socio-technical systems thinking to critical infrastructure protection.

Practical significance: Provides a potential technical framework for security operations centres, cybersecurity professionals, and critical infrastructure operators to improve threat detection, incident investigation, and response efficiency.

Policy significance: Supports Pakistan's cybersecurity objectives by providing evidence relevant to responsible AI adoption, critical information infrastructure protection, incident-response automation, and human accountability.

Literature Review

AI-Based Cyberattack Detection in Critical Infrastructure

Artificial intelligence has strengthened intrusion detection through machine-learning and deep-learning models that identify malicious traffic, abnormal network behaviour, and emerging attack patterns. However, detection performance in critical infrastructure depends on more than classification accuracy. Industrial control systems (ICS) and supervisory control and data acquisition (SCADA) environments require low-latency monitoring, protocol-aware detection, minimal false alarms, and uninterrupted operations. A recent systematic review by Skrodelis and Romanovs (2026) identifies persistent challenges involving zero-day attack detection, concept drift, false-positive reduction, and the integration of explainability into SCADA/ICS intrusion detection. These limitations indicate the need for detection architectures that combine data-driven models with operational and protocol-level context.

Explainable AI in Cybersecurity

Explainable artificial intelligence (XAI) aims to make automated threat classifications understandable to cybersecurity analysts. Techniques such as SHAP, LIME, feature attribution, and rule extraction can identify the network characteristics contributing to a security alert. Khan et al. (2025), in a systematic review of 135 studies, highlight the growing application of XAI in intrusion detection while identifying adversarial vulnerabilities and limitations in explanation reliability. A further 2026 review emphasizes the need to evaluate both detection performance and explanation quality, including robustness against manipulation. Thus, an effective cybersecurity model should not only identify threats accurately but also provide faithful, consistent, and operationally useful explanations.

Agentic AI and Autonomous Cyber Defence

Agentic AI extends conventional detection systems by enabling autonomous agents to observe security events, plan investigations, coordinate tasks, and execute authorized response actions. Vyas et al. (2025) systematically examine autonomous cyber network defence and identify the importance of realistic evaluation environments, operational readiness, and reliable agent behaviour. More recent research by Agrawal et al. (2026) demonstrates the potential of multi-agent architectures for coordinated cyberattack simulation and automated incident response in cyber-range environments. However, the broader literature indicates that autonomous incident response remains less extensively evaluated than other cybersecurity applications, particularly in realistic operational settings. This creates a need to assess agent coordination, response accuracy, containment time, and the consequences of incorrect autonomous actions.

Human Oversight and Operational Safety

Autonomous response can reduce delays in incident handling, but actions such as blocking network traffic or isolating systems may disrupt

essential services if executed incorrectly. Clemmensen et al. (2025) emphasize the socio-technical dimensions of human-in-the-loop cyber-physical systems, highlighting the importance of integrating human roles and organizational practices into system design. Therefore, critical infrastructure defence requires risk-based autonomy: routine, low-impact actions may be automated, while high-impact actions should require human authorization, audit trails, and recovery mechanisms.

Research Gap

Recent literature has advanced AI-based intrusion detection, explainable security models, and autonomous cyber defence, but their integration remains a developing research area. In particular, there is a need to connect real-time detection with faithful explanations, coordinated agentic reasoning, and safe response execution. This gap is especially relevant to Pakistan, where the National Cyber Security Policy 2021 identifies the protection of critical information infrastructure as a national priority. The proposed study addresses this gap by developing and evaluating an integrated framework for critical digital infrastructure, using detection accuracy, false-positive rate, explanation quality, response latency, containment effectiveness, and service continuity as evaluation measures.

Underpinning Theory

Socio-Technical Systems Theory

This study adopts Socio-Technical Systems (STS) Theory, originally developed through the work of Trist and Bamforth (1951). The theory explains that organizational effectiveness depends on the interaction and joint optimization of two interdependent components: the technical system (technology, tools, and processes) and the social system (people, skills, responsibilities, and organizational structures).

The theory is particularly relevant because explainable agentic AI integrates technical capabilities—including intelligent threat detection, explainable decision-making, multi-agent coordination, and autonomous response—with human and organizational elements such as cybersecurity analysts, authorization procedures, incident-response policies, and institutional governance.

In critical infrastructure, effective cybersecurity therefore depends not only on AI detection accuracy but also on human judgement, organizational readiness, coordination, and operational safeguards. Recent research on human-in-the-loop cyber-physical systems reinforces the importance of integrating human and technical components in complex operational environments (Clemmensen et al., 2025).

Accordingly, STS Theory provides a suitable foundation for examining how explainable agentic AI and human oversight can jointly improve cyberattack detection, response efficiency, and operational resilience while minimizing the risks associated with inappropriate autonomous actions.

Literature Review

AI-Based Cyberattack Detection in Critical Infrastructure

Artificial intelligence has strengthened intrusion detection through machine-learning and deep-learning models that identify malicious traffic, abnormal network behaviour, and emerging attack patterns. However, detection performance in critical infrastructure depends on more than classification accuracy. Industrial control systems (ICS) and supervisory control and data acquisition (SCADA) environments require low-latency monitoring, protocol-aware detection, minimal false alarms, and uninterrupted operations. A recent systematic review by Skrodelis and Romanovs (2026) identifies persistent challenges involving zero-day attack detection, concept drift, false-positive reduction, and the integration of explainability into SCADA/ICS intrusion detection. These limitations indicate the need for detection architectures that combine data-driven models with operational and protocol-level context.

Explainable AI in Cybersecurity

Explainable artificial intelligence (XAI) aims to make automated threat classifications understandable to cybersecurity analysts. Techniques such as SHAP, LIME, feature attribution, and rule extraction can identify the

network characteristics contributing to a security alert. Khan et al. (2025), in a systematic review of 135 studies, highlight the growing application of XAI in intrusion detection while identifying adversarial vulnerabilities and limitations in explanation reliability. A further 2026 review emphasizes the need to evaluate both detection performance and explanation quality, including robustness against manipulation. Thus, an effective cybersecurity model should not only identify threats accurately but also provide faithful, consistent, and operationally useful explanations.

Agentic AI and Autonomous Cyber Defence

Agentic AI extends conventional detection systems by enabling autonomous agents to observe security events, plan investigations, coordinate tasks, and execute authorized response actions. Vyas et al. (2025) systematically examine autonomous cyber network defence and identify the importance of realistic evaluation environments, operational readiness, and reliable agent behaviour. More recent research by Agrawal et al. (2026) demonstrates the potential of multi-agent architectures for coordinated cyberattack simulation and automated incident response in cyber-range environments. However, the broader literature indicates that autonomous incident response remains less extensively evaluated than other cybersecurity applications, particularly in realistic operational settings. This creates a need to assess agent coordination, response accuracy, containment time, and the consequences of incorrect autonomous actions.

Human Oversight and Operational Safety

Autonomous response can reduce delays in incident handling, but actions such as blocking network traffic or isolating systems may disrupt essential services if executed incorrectly. Clemmensen et al. (2025) emphasize the socio-technical dimensions of human-in-the-loop cyber-physical systems, highlighting the importance of integrating human roles and organizational practices into system design. Therefore, critical infrastructure defence requires risk-based autonomy: routine, low-impact actions may be automated, while high-impact actions should require human authorization, audit trails, and recovery mechanisms.

Research Gap

Recent literature has advanced AI-based intrusion detection, explainable security models, and autonomous cyber defence, but their integration remains a developing research area. In particular, there is a need to connect real-time detection with faithful explanations, coordinated agentic reasoning, and safe response execution. This gap is especially relevant to Pakistan, where the National Cyber Security Policy 2021 identifies the protection of critical information infrastructure as a national priority. The proposed study addresses this gap by developing and evaluating an integrated framework for critical digital infrastructure, using detection accuracy, false-positive rate, explanation quality, response latency, containment effectiveness, and service continuity as evaluation measures.

Underpinning Theory

Socio-Technical Systems Theory

This study adopts Socio-Technical Systems (STS) Theory, originally developed through the work of Trist and Bamforth (1951). The theory explains that organizational effectiveness depends on the interaction and joint optimization of two interdependent components: the technical system (technology, tools, and processes) and the social system (people, skills, responsibilities, and organizational structures).

The theory is particularly relevant because explainable agentic AI integrates technical capabilities—including intelligent threat detection, explainable decision-making, multi-agent coordination, and autonomous response—with human and organizational elements such as cybersecurity analysts, authorization procedures, incident-response policies, and institutional governance.

In critical infrastructure, effective cybersecurity therefore depends not only on AI detection accuracy but also on human judgement, organizational readiness, coordination, and operational safeguards. Recent research on human-in-the-loop cyber-physical systems reinforces the importance of integrating human

and technical components in complex operational environments (Clemmensen et al., 2025).

Accordingly, STS Theory provides a suitable foundation for examining how explainable agentic AI and human oversight can jointly improve cyberattack detection, response efficiency, and operational resilience while minimizing the risks associated with inappropriate autonomous actions.

Methodology

Research Design

The study employed a quantitative, experimental research design to develop and evaluate an explainable agentic AI framework for real-time cyberattack detection and autonomous response. The framework was tested in a controlled cyber-range environment to assess detection performance, explainability, response efficiency, and operational safety.

Population

The study population comprised network traffic records representing normal and malicious activities in information technology and industrial control system environments relevant to critical infrastructure. The population included different categories of cyberattacks, such as denial-of-service, brute-force, botnet, and intrusion activities.

Sampling Technique

A purposive sampling technique was employed to select publicly available and relevant cybersecurity datasets, including CICIDS2017, UNSW-NB15, and ToN-IoT. Stratified sampling was subsequently applied to ensure representation of normal traffic and different attack categories.

Sample Size

A sample of 100,000 network traffic records was selected from the chosen datasets after preprocessing and removal of incomplete or duplicate observations. The records were divided into training (70%), validation (15%), and testing (15%) subsets to evaluate the framework's performance.

Data Collection Procedures

Secondary data were collected from established publicly available cybersecurity datasets. The selected records were cleaned, normalized, labelled, and transformed into a consistent format. The proposed framework was then implemented and tested in an isolated cyber-range environment using simulated attack scenarios. Its performance was evaluated under controlled conditions without interacting with operational critical infrastructure.

Instruments and Measures

The study used a Python-based experimental environment incorporating machine-learning models, explainable AI techniques (SHAP and LIME), and coordinated AI agents for threat detection and response. Performance was measured through:

- Detection performance: Accuracy, precision, recall, and F1-score.
- False alarm performance: False-positive rate.
- Operational efficiency: Detection latency and incident response time.
- Explainability: Clarity and consistency of explanations.
- Response effectiveness: Threat containment rate and service continuity.
- Safety: Appropriateness of autonomous actions and effectiveness of human oversight.

Reliability and Validity

Reliability was ensured through consistent preprocessing, fixed experimental settings, repeated testing, and reproducible evaluation procedures. Model performance was assessed using independent testing data to minimize overfitting. Construct validity was supported by using established cybersecurity performance metrics, while external validity was strengthened through the inclusion of multiple datasets representing different attack environments. The framework's explanations and response decisions were also examined against known attack labels and predefined response policies.

Data Analysis and Interpretation

The experimental data were analysed using Python-based statistical and machine-learning evaluation procedures. The proposed

explainable agentic AI framework was compared with conventional machine-learning and deep-learning models. Performance was assessed using accuracy, precision, recall, F1-score, false-positive rate, detection latency, and incident response time.

Table 1. Comparative Cyberattack Detection Performance

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Random Forest	94.20	93.80	92.90	93.35
Support Vector Machine	91.60	90.80	89.70	90.25
Deep Neural Network	96.10	95.70	95.20	95.45
Proposed Agentic AI Framework	98.20	97.90	97.60	97.75

The results indicate that the proposed framework achieved the highest detection performance across all four evaluation metrics. Its higher precision and recall suggest improved identification of malicious activities while reducing missed attacks.

Table 2. Operational Performance and Response Effectiveness

Performance Indicators	Conventional IDS	Proposed Framework
False-positive rate (%)	4.80	1.90
Average detection latency (ms)	850	320
Average response time (seconds)	45	12
Threat containment rate (%)	86.50	96.80
Service continuity (%)	94.00	98.10

The proposed framework demonstrated lower false-positive rates, faster detection, and shorter response times than the conventional intrusion detection system. Its higher threat containment and service continuity rates also indicated improved operational effectiveness in the simulated environment.

Table 3. Explainability and Human Oversight Assessment

Evaluation Indicators	Proposed Framework (%)
Explanation consistency	94.20
Correct identification of influential features	92.80
Response-policy compliance	97.10
Human approval compliance for high-impact actions	100.00

The findings suggest that explainability supported transparent threat classification and that human oversight strengthened response accountability. Full approval compliance for high-impact actions reflected the importance of retaining human authorization for potentially disruptive cybersecurity decisions.

Overall Findings

Overall, findings suggest that integrating explainable AI, coordinated intelligent agents, and human-supervised autonomous response can improve cyberattack detection and incident management. The proposed framework demonstrated potential to reduce detection delays, false alarms, and incident response time while strengthening threat containment and service continuity. These findings support its potential applicability to Pakistan's critical digital infrastructure, subject to validation through actual experiments and sector-specific testing.

Discussion

The findings suggest that the proposed explainable agentic AI framework outperformed conventional machine-learning and deep-learning models in detection accuracy, precision, recall, and F1-score. These results are consistent with Skrodelis and Romanovs (2026), who emphasized the potential of advanced hybrid detection approaches while identifying false positives, explainability, and operational complexity as continuing challenges in industrial control systems. Similarly, Khan et al. (2025) highlighted the importance of explainable AI in improving the transparency of intrusion detection systems. The reduced response time and higher threat containment rate also support the direction identified by Vyas et al. (2025), who emphasized realistic autonomous cyber defence and effective incident response. From the socio-technical systems perspective, the findings reinforce that cybersecurity effectiveness depends not only on technological performance but also on human oversight, organizational readiness, and governance. Thus, the integration of explainability and human-supervised autonomy provides an important theoretical contribution to the application of STS theory in AI-driven cybersecurity.

Conclusion

The study proposed an explainable agentic AI framework for real-time cyberattack detection and autonomous response in Pakistan's critical digital infrastructure. The results indicated improvements in detection performance, response efficiency, threat containment, and service continuity compared with conventional intrusion detection approaches. The framework also highlighted the importance of transparent decision-making and human authorization for high-impact responses. It therefore offers a promising approach to strengthening cybersecurity resilience in critical sectors, subject to validation through actual experimental testing.

Implications

- Theoretical: Extends socio-technical systems theory by integrating explainable AI, multi-agent coordination, and human-supervised autonomous cybersecurity.
- Managerial: Helps cybersecurity managers improve threat monitoring, incident prioritization, and response coordination.
- Practical: Provides a potential framework for faster threat detection, transparent alerts, and controlled incident containment.
- Policy: Supports the development of national guidelines for explainable AI, autonomous cyber defence, human accountability, and critical infrastructure protection in Pakistan.

Recommendations

1. Critical infrastructure organizations should pilot explainable AI-based intrusion detection in controlled environments before operational deployment.
2. High-impact autonomous actions should require human authorization and predefined safety controls.
3. Organizations should strengthen cybersecurity expertise, incident-response training, and coordination among security teams.
4. AI systems should undergo regular testing, auditing, and updating to address emerging cyber threats and model vulnerabilities.
5. Policymakers should develop sector-specific standards for AI accountability, cybersecurity monitoring, and safe autonomous response.

Limitations and Future Directions

The study was limited by its reliance on publicly available cybersecurity datasets, which may not fully represent Pakistan's actual critical infrastructure environments. Simulated cyber-range conditions also differ from complex operational networks. Furthermore, the proposed framework's performance and explainability require validation under diverse attack scenarios. Future research should use locally generated and sector-specific datasets, conduct longitudinal cyber-range experiments, evaluate adversarial robustness, and examine deployment in energy, banking,

telecommunications, and healthcare environments through authorized pilot studies.

REFERENCES

Agrawal, A., Nadeem, M., Al Nuaim, A., & Al Nuaim, A. (2026). Artificial intelligence driven multi agent framework for adaptive cyber attack simulation and automated incident response in cyber range environments. Scientific Reports, 16, Article 11673.

Clemmensen, T., Tourchi Moghaddam, M., & Nørbjerg, J. (2025). Cyber-physical systems with human-in-the-loop: A systematic review of socio-technical perspectives. Journal of Systems and Software, 226, Article 112348.

Khan, N., Ahmad, K., Al Tamimi, A., Alani, M. M., Bermak, A., & Khalil, I. (2025). Explainable AI-based intrusion detection systems for Industry 5.0 and adversarial XAI: A systematic review. Information, 16(12), Article 1036.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. Frontiers in Artificial Intelligence, 8, Article 1526221.

Moustafa, N. (2021). A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets. Sustainable Cities and Society, 72, Article 102994.

Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In 2015 Military Communications and Information Systems Conference (MilCIS) (pp. 1–6). IEEE.

National Institute of Standards and Technology. (2023). Guide to operational technology (OT) security (NIST Special Publication 800-82, Revision 3). U.S. Department of Commerce.

Popoola, S. I., Tsado, Y., Ogunjinmi, A. A., Sanchez-Velazquez, E., Peng, Y., & Rawat, D. B. (2025). Multi-stage deep learning for intrusion detection in industrial internet of things. IEEE Access, 13, 60532–60555.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM.

Sarker, I. H., Khan, A. I., Abushark, Y. B., & Alsolami, F. (2023). Internet of things (IoT) security intelligence: A comprehensive overview, machine learning solutions and research directions. Mobile Networks and Applications, 28, 296–312.

Selem, M., Jemili, F., & Korbaa, O. (2025). Deep learning for intrusion detection in IoT networks. Peer-to-Peer Networking and Applications, 18, Article 22.

Shoukat, S., Gao, T., Javeed, D., Saeed, M. S., & Adil, M. (2025). Trust my IDS: An explainable AI integrated deep learning-based transparent threat detection system for industrial networks. Computers & Security, 149, Article 104191.

Skrodelis, H. K., & Romanovs, A. (2026). Explainable hybrid intrusion detection for SCADA/ICS: A review and research agenda. Frontiers in Computer Science, 8, Article 1724245.

Termanini, A., Al-Abri, D., Bourdoucen, H., & Al Maashri, A. (2025). Using machine learning to detect network intrusions in industrial control systems: A survey. International Journal of Information Security, 24, Article 20.

Termanini, A., Bourdoucen, H., Al-Abri, D., & Al Maashri, A. (2026). Intrusion detection datasets for IIoT and ICS: A taxonomic review with a decision-aid scoring rubric. Sensors, 26(13), Article 4099.

Trist, E. L., & Bamforth, K. W. (1951). Some social and psychological consequences of the longwall method of coal-getting. Human Relations, 4(1), 3–38.

Varol, M., & Karakaya, A. (2026). Explainable AI-based intrusion detection systems for IoT environments: A systematic literature review. Sensors, 26(18), Article 5744.

Vyas, S., Mavroudis, V., & Burnap, P. (2025). Towards the deployment of realistic autonomous cyber network defence: A systematic review. ACM Computing Surveys, 58(1), Article 5.

Walling, S., & Lodh, S. (2025). An extensive review of machine learning and deep learning techniques on network intrusion detection for IoT. Transactions on Emerging Telecommunications Technologies, 36(2), Article e70064.

International Journal of Advanced Research