

EXPLAINABLE FEDERATED ARTIFICIAL INTELLIGENCE FOR ZERO-DAY CYBERATTACK DETECTION IN PAKISTAN'S CRITICAL DIGITAL INFRASTRUCTURE

Iqra Iqbal^{*1}, Dr Eram Abbasi², Dr Hina Gul³

^{*1}Department Computer Science University, University of Peshwar

²Asst Prof Computer Science IQRA University

³Assistant professor computer science PhD computer science HED, Government Girls Degree College Matta

²eramabbasi@gmail.com, ³hinagul.ggdc@gmail.com

DOI: <https://doi.org/10.5281/zenodo.21898792>

Keywords

Explainable Artificial Intelligence; Federated Learning; Zero-Day Cyberattacks; Intrusion Detection; Critical Digital Infrastructure; Cybersecurity

Article History

Received: 01 October 2025

Accepted: 16 December 2025

Published: 31 December 2025

Copyright @Author

Corresponding Author: *

Iqra Iqbal

Abstract

The increasing sophistication of cyberattacks poses significant risks to Pakistan's critical digital infrastructure, while conventional intrusion detection systems face limitations in zero-day detection, data centralization, and interpretability. This study evaluates an Explainable Federated Artificial Intelligence (XAI-FL) framework for privacy-preserving zero-day cyberattack detection across distributed critical-infrastructure environments. A quantitative experimental design was used with cybersecurity network-flow data distributed among simulated organizational clients. XAI-FL was compared with centralized artificial intelligence and conventional federated learning using accuracy, precision, recall, F1-score, AUROC, false-positive rate, and false-negative rate. Additional experiments examined zero-day detection, non-IID data heterogeneity, explainability, and malicious client participation. XAI-FL achieved superior overall performance and stronger generalization to unseen attacks while providing feature-level explanations. Performance declined as data heterogeneity increased, whereas robust aggregation reduced the effect of malicious client updates. The findings suggest that integrating XAI with federated learning provides a promising foundation for privacy-aware, interpretable, and collaborative zero-day detection. However, real-world validation using Pakistani critical-infrastructure data is required.

1. INTRODUCTION

The increasing digitalization of critical infrastructure has expanded dependence on interconnected information and operational technologies while increasing exposure to sophisticated cyber threats. Energy, telecommunications, finance, healthcare, transportation, and government systems are particularly vulnerable to attacks that exploit previously unknown vulnerabilities. Zero-day attacks are difficult to detect because conventional

signature-based systems depend on previously identified attack characteristics (Dai et al., 2024). Artificial intelligence (AI), particularly machine learning and deep learning, has improved intrusion detection by identifying behavioral patterns associated with malicious activity. However, strong benchmark performance does not necessarily translate into reliable deployment in heterogeneous real-world environments. Centralized AI systems also require organizations to transfer sensitive cybersecurity data to common

repositories. Federated learning (FL) addresses this limitation by allowing organizations to retain data locally while sharing model updates. Nevertheless, FL introduces challenges related to non-IID data, communication costs, malicious clients, and model security (Belenguer et al., 2025; Zhang et al., 2025).

Another limitation of AI-based intrusion detection is poor interpretability. Explainable artificial intelligence (XAI) provides information about the features contributing to model predictions, potentially improving analyst trust and decision-making (Oki et al., 2024). This is particularly important for zero-day detection, where analysts may lack predefined signatures.

Existing research has examined zero-day detection, federated intrusion detection, XAI, and adversarial robustness separately. However, limited empirical research integrates these dimensions within a single framework, particularly in Pakistan's critical digital infrastructure.

Therefore, this study proposes and evaluates an XAI-FL framework for zero-day cyberattack detection. It examines detection performance, zero-day generalization, the effects of heterogeneous data, model interpretability, and resilience against malicious federated clients.

2. Problem Statement

Critical digital infrastructure increasingly depends on interconnected technologies, creating exposure to sophisticated cyberattacks. Conventional signature-based systems are limited in detecting previously unknown attacks, while centralized AI approaches raise concerns regarding data privacy, governance, and security.

Federated learning can preserve local data ownership but introduces challenges related to non-IID distributions, communication overhead, malicious clients, and model poisoning. At the same time, complex AI models may provide accurate predictions without sufficient explanations for cybersecurity analysts.

The central problem is therefore the absence of a unified, privacy-preserving, explainable, and robust AI framework capable of detecting zero-day attacks across heterogeneous critical-infrastructure

environments without requiring centralized access to sensitive cybersecurity data.

3. Research Questions

RQ1: How effectively can XAI-FL detect zero-day cyberattacks in heterogeneous critical-infrastructure environments?

RQ2: How does XAI-FL compare with centralized and conventional federated intrusion-detection models?

RQ3: How does non-IID data heterogeneity affect federated zero-day detection performance?

RQ4: To what extent does XAI improve the interpretability of zero-day predictions?

RQ5: How resilient is XAI-FL to malicious or manipulated client updates?

RQ6: What computational and communication costs are associated with XAI-FL deployment?

4. Research Objectives

General Objective

To develop and empirically evaluate an Explainable Federated Artificial Intelligence framework for privacy-preserving zero-day cyberattack detection in Pakistan's critical digital infrastructure.

Specific Objectives

1. To develop a federated AI framework for detecting previously unseen attacks.
2. To evaluate detection performance using accuracy, precision, recall, F1-score, AUROC, FPR, and FNR.
3. To compare XAI-FL with centralized and conventional federated approaches.
4. To examine the effects of non-IID data distributions.
5. To evaluate the interpretability of XAI-based predictions.
6. To assess resilience against malicious client updates.
7. To examine computational and communication overhead.
8. To develop recommendations for adopting and governing XAI-FL in Pakistan.

5. Significance of the Study

The study integrates four major cybersecurity requirements: zero-day detection, federated learning, explainability, and adversarial robustness. It extends existing research by examining these dimensions within a unified framework.

Practically, XAI-FL can enable organizations to collaborate on cyberattack detection without routinely transferring raw cybersecurity data. XAI can also help analysts interpret suspicious activity and prioritize security alerts.

For managers, the findings highlight the importance of considering detection performance, privacy, data heterogeneity, model security, computational requirements, and human oversight rather than accuracy alone.

For policymakers, the study supports privacy-preserving cybersecurity collaboration, secure federated learning, explainability, AI auditing, incident reporting, and accountability in critical-infrastructure protection.

6. Literature Review

6.1 AI-Based Zero-Day Detection

Zero-day attacks are difficult to detect because their characteristics may not have been previously identified. Recent machine-learning and anomaly-detection approaches demonstrate potential for detecting unseen attacks, but benchmark accuracy alone does not establish genuine zero-day capability. Robust evaluation requires deliberately withholding attack categories from training (Dai et al., 2024; Yee Por et al., 2024).

6.2 Federated Learning for Intrusion Detection

Federated learning allows organizations to train models collaboratively without sharing raw data. It is therefore relevant to privacy-sensitive cybersecurity environments. However, FL introduces challenges involving communication, data distributions, aggregation, and malicious participants (Hernández-Ramos et al., 2023; Nguyen & Beuran, 2024).

6.3 Non-IID Data

Different organizations generate different network traffic and experience different attack patterns.

Such non-IID distributions can reduce global model performance. Research therefore emphasizes client selection, adaptive aggregation, and other approaches to manage heterogeneity (Djaidja et al., 2024; Neto et al., 2024).

6.4 Explainable AI in Cybersecurity

XAI techniques such as SHAP and LIME can identify features contributing to model predictions. This can improve transparency and support analysts in investigating suspicious activity. However, explanations do not necessarily establish causality and may vary across models and datasets (Oki et al., 2024; Subasi et al., 2024).

6.5 Explainable Federated Learning

Combining XAI and FL creates additional challenges because federated models incorporate information from clients with potentially different data distributions. Important questions concern explanation stability, client-specific feature importance, and the relationship between local and global explanations. Existing research remains limited regarding XAI for zero-day detection under heterogeneous federated conditions.

6.6 Security of Federated Learning

Federated systems can be attacked through malicious or compromised clients that manipulate model updates. Robust aggregation, client validation, participant selection, and update monitoring are therefore important for secure federated intrusion detection.

6.7 Critical Digital Infrastructure

Critical infrastructure requires cybersecurity systems that provide accuracy, privacy, reliability, interpretability, scalability, and operational continuity. Consequently, evaluation should extend beyond accuracy to include robustness, communication overhead, explainability, and operational reliability.

6.8 Research Gap

Existing studies generally address subsets of the following dimensions:

Zero-day detection + Federated learning +
Explainability + Heterogeneous data +

Adversarial robustness + Pakistan-specific critical infrastructure

Limited research integrates all these dimensions within a single empirical framework. This study addresses this gap through an XAI-FL approach.

7. Hypotheses

H1: XAI-FL significantly improves zero-day detection performance compared with centralized intrusion detection.

H2: Federated AI improves zero-day detection under distributed data conditions compared with conventional non-collaborative approaches.

H3: Higher non-IID data heterogeneity negatively affects zero-day detection performance.

H4: XAI improves the interpretability of federated zero-day predictions.

H5: Robust aggregation improves resilience against malicious client updates.

H6: Greater client heterogeneity increases computational and communication costs.

8.3 Sample

The study used **100,000 network-flow observations**, distributed as:

Category	Frequency	Percentage
Benign traffic	50,000	50%
Known attacks	35,000	35%
Zero-day attacks	15,000	15%
Total	100,000	100%

The observations were distributed among **10 simulated federated clients**, with approximately 10,000 observations per client. The data were divided into training, validation, and testing sets using a 70%, 15%, and 15% allocation.

Important: The original manuscript explicitly states that these numerical values are study values and should be replaced with the actual dataset and experimental figures before submission.

8.4 Data Collection and Processing

The procedure involved:

1. Dataset identification and screening.
2. Data extraction.
3. Cleaning of missing, duplicate, and invalid records.

8. Methodology

8.1 Research Design

A quantitative, experimental, and comparative design was employed. Three configurations were evaluated:

1. Centralized AI
2. Conventional federated learning
3. Proposed XAI-FL

Models were evaluated under IID and non-IID conditions, as well as zero-day and malicious-client scenarios.

8.2 Population and Sampling

The target population consisted of cybersecurity observations relevant to Pakistan's critical infrastructure. Because operational Pakistani network data were unavailable, publicly available cybersecurity datasets were used.

Purposive sampling was applied to select relevant datasets, followed by stratified sampling to maintain representation of benign and malicious traffic.

4. Feature transformation and normalization.
5. Training, validation, and testing partitioning.
6. Withholding selected attack categories for zero-day evaluation.
7. Distribution among simulated federated clients.
8. Local federated model training.
9. XAI-based feature attribution.
10. Malicious-client and robust-aggregation testing.

8.5 Measures

Performance was assessed using:

- Accuracy

- Precision
- Recall
- F1-score
- AUROC
- False-positive rate
- False-negative rate
- Training and inference time
- Communication overhead
- Convergence behavior

SHAP and/or LIME were used for explainability, while FedAvg served as the baseline aggregation method.

8.6 Reliability and Validity

Reliability was assessed through repeated experimental procedures, consistent preprocessing, standardized partitions, model configurations, convergence monitoring, and client-level comparisons.

Construct, internal, external, and ecological validity were addressed through the experimental

design. However, generalization to operational Pakistani infrastructure remains limited because publicly available datasets cannot fully represent live national infrastructure.

8.7 Ethical Considerations

The study used publicly available cybersecurity data and did not collect personally identifiable information. Data were processed in controlled environments, and experimental attacks were conducted only in research settings rather than operational systems.

9. Data Analysis and Results

9.1 Analytical Sample

The sample consisted of 50% benign traffic, 35% known attacks, and 15% unseen attacks. The separate zero-day category allowed evaluation of generalization to attacks excluded from training.

9.2 Comparative Model Performance

Model	Accuracy	Precision	Recall	F1	AUROC
Centralized AI	95.1%	94.3%	92.8%	93.5%	0.963
Conventional FL	93.8%	92.7%	90.9%	91.8%	0.951
XAI-FL	97.2%	96.4%	95.8%	96.1%	0.978

XAI-FL achieved the strongest overall performance, with 97.2% accuracy, 95.8% recall, 96.1% F1-score, and 0.978 AUROC.

9.3 Zero-Day Detection

Model	Accuracy	Precision	Recall	F1	FNR
Centralized AI	89.6%	87.9%	84.7%	86.3%	15.3%
Conventional FL	87.9%	85.8%	82.6%	84.2%	17.4%
XAI-FL	93.4%	91.8%	90.7%	91.2%	9.3%

XAI-FL maintained the highest performance on unseen attacks, achieving 90.7% recall and a 9.3% false-negative rate.

9.4 Effect of Data Heterogeneity

Distribution	Accuracy	Precision	Recall	F1
IID	97.2%	96.4%	95.8%	96.1%
Moderate non-IID	95.8%	94.7%	93.5%	94.1%
High non-IID	93.6%	92.5%	90.8%	91.6%

Performance declined as client heterogeneity increased, with F1 decreasing from 96.1% under

IID conditions to 91.6% under highly non-IID conditions.

9.5 Explainability Analysis

The most influential features were:

Rank	Feature	Importance
1	Flow duration	18.7%
2	Packet transmission rate	16.4%
3	Destination-port behavior	14.9%
4	Connection frequency	12.8%
5	Packet-size variation	10.6%

The findings indicate that these features contributed substantially to model predictions. However, feature attribution explains model

decisions and should not be interpreted as evidence of causal relationships.

9.6 Robustness Against Malicious Clients

Malicious Clients	Standard Aggregation F1	Robust Aggregation F1
0%	96.1%	96.1%
10%	92.8%	95.3%
20%	87.4%	93.7%
30%	81.6%	90.9%

At 30% malicious clients, robust aggregation maintained an F1-score of 90.9%, compared with 81.6% under standard aggregation.

9.7 False-Positive and False-Negative Rates

Model	FPR	FNR
Centralized AI	4.7%	7.2%
Conventional FL	6.2%	9.1%
XAI-FL	3.6%	4.2%

XAI-FL produced the lowest false-positive and false-negative rates among the three models.

9.8 Statistical Analysis

The original study reports significant differences among model configurations, including a 4.3-percentage-point F1 difference between XAI-FL and conventional FL. However, the manuscript explicitly identifies these statistical values as provisional and states that actual repeated-run results must replace them before submission.

9.9 Hypothesis Testing

The reported results support H1–H5 and provisionally support H6. However, the manuscript states that final hypothesis decisions should be based on actual repeated experimental runs and appropriate inferential tests rather than descriptive percentages alone.

10. Discussion

The findings indicate that XAI-FL provides a promising approach to zero-day detection in distributed critical-infrastructure environments. It achieved stronger overall and zero-day performance than centralized AI and conventional FL.

The zero-day results demonstrate that performance decreases when models encounter attacks excluded from training, but XAI-FL retained the strongest performance. This supports the potential of combining behavioral learning with federated architectures while recognizing that benchmark-based zero-day testing does not establish universal protection against genuine real-world zero-day exploits.

Conventional FL performed below centralized AI, reflecting challenges associated with distributed and heterogeneous data. However, centralized learning requires data pooling, whereas FL provides privacy and data-sovereignty advantages that may be more important in critical infrastructure.

Performance also declined as non-IID heterogeneity increased. This demonstrates the need for personalized, clustered, or adaptive federated-learning approaches in environments where organizations have substantially different network characteristics.

The explainability results indicate that network characteristics such as flow duration, packet transmission rate, destination-port behavior, and

connection frequency contributed substantially to predictions. XAI can therefore support analyst interpretation, although explanations should be treated as decision-support rather than causal evidence.

Finally, malicious-client experiments demonstrate that federated learning itself can become an attack surface. Robust aggregation substantially reduced performance degradation, highlighting the need for secure aggregation, client authentication, update validation, and Byzantine-resilient mechanisms.

Theoretical Perspective

From a Socio-Technical Systems perspective, successful XAI-FL deployment depends not only on technical performance but also on cybersecurity professionals, organizational structures, data governance, institutional trust, and human oversight.

11. Conclusion

This study evaluated an XAI-FL framework for zero-day cyberattack detection in Pakistan's critical digital infrastructure. XAI-FL demonstrated stronger overall and zero-day detection performance than centralized AI and conventional FL while providing interpretable feature-level information.

Detection performance declined under increasing data heterogeneity, confirming non-IID distributions as a major challenge. Robust aggregation reduced the impact of malicious federated clients, demonstrating the importance of securing the collaborative learning process.

Overall, XAI-FL provides a promising foundation for privacy-aware, interpretable, and collaborative cyber defense. However, deployment in Pakistan requires real-world validation, secure federated governance, heterogeneous-data management, continuous monitoring, and human oversight.

12. Implications

12.1 Theoretical Implications

The study integrates zero-day detection, federated learning, explainability, and adversarial robustness into a unified cybersecurity framework. It further supports a socio-technical perspective in which

effective AI deployment depends on both technical and organizational factors.

12.2 Managerial Implications

Managers should evaluate not only predictive accuracy but also zero-day detection, false negatives, interpretability, privacy, heterogeneity, robustness, computational requirements, and analyst acceptance. Human-in-the-loop procedures should be maintained for high-impact cybersecurity decisions.

12.3 Practical Implications

XAI-FL can enable collaborative AI training while keeping sensitive data locally. XAI may assist security operations centers with alert prioritization, incident investigation, threat hunting, and analyst decision-making.

12.4 Policy Implications

Pakistan's cybersecurity policies should support privacy-preserving collaboration, secure information sharing, AI-enabled intrusion detection, explainability and auditability, secure federated learning, malicious-update detection, incident reporting, and human accountability.

13. Recommendations

1. Develop sector-based federated cybersecurity networks.
2. Evaluate AI models using genuinely unseen and temporally separated attacks.
3. Adopt robust aggregation and client-validation mechanisms.
4. Integrate XAI into security operations centers.
5. Address non-IID data through personalized or clustered FL.
6. Establish continuous model-performance monitoring.
7. Maintain human oversight for high-risk cybersecurity decisions.
8. Conduct pilot projects with Pakistani critical-infrastructure organizations.
9. Develop national guidance for AI-based cyber defense.

14. Limitations and Future Research

The study has several limitations. First, it relies on publicly available datasets rather than operational Pakistani critical-infrastructure data. Second, zero-day attacks were simulated by withholding selected attack categories rather than observing genuine unknown vulnerabilities. Third, federated clients were simulated rather than represented by independent operational organizations. Fourth, some reported numerical and statistical results are provisional and require replacement with actual experimental outputs. Finally, the adversarial analysis covered selected malicious-client scenarios, while more sophisticated attacks may produce different results.

Future research should focus on:

- Real-world Pakistani cybersecurity datasets
- Cross-sector federated learning
- Advanced zero-day and anomaly-detection methods
- Personalized federated learning
- Byzantine-resilient and privacy-preserving mechanisms
- Human-centered XAI evaluation
- Real-time edge deployment
- Pakistan-specific national cybersecurity architecture.

15. Final Research Perspective

Zero-day cyberattack detection in critical infrastructure should be viewed not merely as a classification problem but as a combined challenge involving privacy, distributed intelligence, interpretability, **robustness**, organizational collaboration, and governance. The proposed XAI-FL framework provides a multidimensional foundation for future cybersecurity research in Pakistan, but its practical value ultimately depends on secure, transparent, efficient, and responsible deployment in real critical-infrastructure environments.

Important editorial note

The original manuscript is 40 pages, and a substantial amount of repetition has now been removed. The biggest issue I would address before journal submission is not length but the

manuscript's own warning that several numerical/statistical results are provisional—for example, the ANOVA results and hypothesis decisions.

Also, RQ6 concerns computational and communication costs, but the presented Results section does not provide a corresponding numerical analysis of those costs. That should either be added with actual results or RQ6/objective 7 should be removed to maintain consistency. This is an important methodological/editorial issue rather than merely a shortening issue.

REFERENCES

- Agrawal, S., Sarkar, S., Aouedi, O., Yenduri, G., Piamrat, K., Bhattacharya, S., Maddikunta, P. K. R., & Gadekallu, T. R. (2022). Federated learning for intrusion detection system: Concepts, challenges and future directions. *Computer Communications*, 195, 346–361.
- Ali, A., Ullah, M., Khan, M. T., & Shehzad, U. (2026). Impact of artificial intelligence-based predictive analytics on improving academic performance in Pakistani universities: The moderating role of digital literacy. *Spectrum of Engineering Sciences*, 4(3), 167–178.
- Buyuktanir, B., Altinkaya, Ş., Karatas Baydogmus, G., & Yildiz, K. (2025). Federated learning in intrusion detection: Advancements, applications, and future directions. *Cluster Computing*, 28, Article 473.
- Dai, Z., Por, L. Y., Chen, Y.-L., Yang, J., Ku, C. S., Alizadehsani, R., & Pławiak, P. (2024). An intrusion detection model to detect zero-day attacks in unseen data using machine learning. *PLOS ONE*, 19(9), Article e0308469.
- Djaidja, T. E. T., Brik, B., Boualouache, A., Senouci, S. M., & Ghamri-Doudane, Y. (2024). Federated learning for 5G and beyond: A blessing and a curse—An experimental study on intrusion detection systems. *Computers & Security*, 139, Article 103707.
- Dos Santos, R. R., Viegas, E. K., Santin, A. O., & Tedeschi, P. (2023). Federated learning for reliable model updates in network-based intrusion detection. *Computers & Security*, 133, Article 103413.
- Hernández-Ramos, J. L., Karopoulos, G., Chatzoglou, E., Kouliaridis, V., Mármol, E., Gonzalez-Vidal, A., & Kambourakis, G. (2023). Intrusion detection based on federated learning: A systematic review. *Computer Networks*, 241, Article 110214.
- Hossain, M. A., Saif, S., & Islam, M. S. (2025). A novel federated learning approach for IoT botnet intrusion detection using SHAP-based knowledge distillation. *Complex & Intelligent Systems*, 11, Article 422.
- Javeed, D., Saeed, M. S., Adil, M., Kumar, P., & Jolfaei, A. (2024). A federated learning-based zero trust intrusion detection system for Internet of Things. *Ad Hoc Networks*, 162, Article 103540.
- Khan, N., Ahmad, K., Al Tamimi, A., Alani, M. M., Bermak, A., & Khalil, I. (2024). Explainable AI-based intrusion detection system for Industry 5.0: An overview of the literature, associated challenges, the existing solutions, and potential research directions. *arXiv*.
- Khraisat, A., Alazab, A., Alazab, M., Obeidat, A., Singh, S., & Jan, T. (2025). Federated learning for intrusion detection in IoT environments: A privacy-preserving strategy. *Discover Internet of Things*, 5, Article 72.
- Lai, Y. C., Lin, J. Y., Lin, Y. D., Hwang, R. H., Lin, P. C., Wu, H. K., & Chen, C. K. (2023). Two-phase defense against poisoning attacks on federated learning-based intrusion detection. *Computers & Security*, 129, Article 103205.
- Latif, N., Ma, W., & Ahmad, H. B. (2025). Advancements in securing federated learning with IDS: A comprehensive review of neural networks and feature engineering techniques for malicious client detection. *Artificial Intelligence Review*, 58, Article 91.

- Liang, Y., & Luo, M. (2025). Optimization of distributed network intrusion detection system based on Internet of Things and federated learning. *Discover Internet of Things*, 5, Article 60.
- Mármol Campos, E., Gonzalez-Vidal, A., Hernández-Ramos, J. L., & Skarmeta, A. (2025). Federated learning for misbehaviour detection with variational autoencoders and Gaussian mixture models. *International Journal of Information Security*, 24, Article 95.
- Qazi, S., Ullah, M., Khalil, Y. K., & Iqbal, S. (2026). Fintech adoption and financial inclusion in Pakistan: The role of digital payment platforms in enhancing access to formal financial services. *International Journal of Social Sciences Bulletin*, 4(3), 718–732.
- Sardar, H., Farooq, S. U., Ullah, M., & Habib, A. B. (2025). Impact of digital financial services on poverty alleviation and income inequality in rural Pakistan: Evidence from mobile banking and fintech platforms. *Advance Journal of Econometrics and Finance*, 3(1), 45–62.
- Sarhan, M., Layeghy, S., Gallagher, M., & Portmann, M. (2023). From zero-shot machine learning to zero-day attack detection. *International Journal of Information Security*, 22, 947–959.
- Selvam, P., Karthikeyan, P., Manochitra, S., Sujith, A. V. L. N., Ganesan, T., Ayyasamy, R., Shuaib, M., Alam, S., & Rajendran, A. (2025). Federated learning-based hybrid convolutional recurrent neural network for multi-class intrusion detection in IoT networks. *Discover Internet of Things*, 5, Article 39.
- Shao, J.-M., Zeng, G.-Q., Lu, K.-D., Geng, G.-G., & Weng, J. (2024). Automated federated learning for intrusion detection of industrial control systems based on evolutionary neural architecture search. *Computers & Security*, 143, Article 103910.
- Sharma, B., Sharma, L., Lal, C., & Roy, S. (2024). Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach. *Expert Systems with Applications*, 238, Article 121751.
- Sharma, K. P., Nagpal, T., Vora, T., Yadav, A., Abdullah, M. I., Jayaprakash, B., Kashyap, A., Sridevi, G., Bhowmik, A., & Bukate, B. B. (2025). Interpretable intrusion detection for IoT environments using a self-attention-based explainable AI framework. *Scientific Reports*, 15, Article 39937.
- Vyas, A., Lin, P.-C., Hwang, R.-H., & Tripathi, M. (2024). Privacy-preserving federated learning for intrusion detection in IoT environments: A survey. *IEEE Access*, 12, 127018–127050.
- Wang, J., Yang, T., Chen, W., Deng, H., Huang, Q., & Li, D. (2025). Federated learning with self-updating server parameters for smart home intrusion detection in non-IID environments. *Expert Systems with Applications*, 267, Article 126233.
- Zhang, Y., Song, J., Sun, Y., Gao, Z., Hu, Z., & Sun, M. (2025). Federated two-stage transformer-based network for intrusion detection in non-IID data of controller area networks. *Cybersecurity*, 8, Article 29.