Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

AN EXPLAINABLE FEDERATED DEEP LEARNING FRAMEWORK FOR REAL-TIME INTRUSION DETECTION IN CRITICAL INFRASTRUCTURE IOT NETWORKS

Muhammad Umar[1], Azhar Ali Khan[*2], Ayesha Batool[3], Mustajib Ur Rehman[4], Imran Khursheed[5], Muhammad Abrar[6]

[1]Department of Software Engineering, NUML, Multan, Pakistan.
*[*2]Department of Engineering & Computing, NUML, Multan, Pakistan.*
[3]Department of SE & IT, Federal Urdu University of Arts and Technology, Science, Islamabad, Pakistan
[4]Department of Software Engineering, NUML Multan, Pakistan.
[5]Department of Software Engineering, NUML, Multan, Pakistan
[6]Department of Computer Science, USP, Multan, Pakistan

Corresponding Author: [*2]aakmcs2018@gmail.com

DOI: https://doi.org/10.5281/zenodo.21867963

Keywords
Federated learning, Explainable artificial intelligence, Intrusion detection, Internet of Things, Critical infrastructure, Deep learning, Convolutional neural networks, Cybersecurity, Real-time systems, Edge computing.

Article History
Received: 13 January 2026
Accepted: 26 February 2026
Published: 12 March 2026

Corresponding Author: *
Azhar Ali Khan

Abstract

The paper introduces a novel Explainable Federated Deep Learning Framework for real-time Intrusion Detection in Critical Infrastructure IoT Networks, to offer secure, scalable, privacy-preserving, and interpretable cybersecurity. With the rapid growth of IoT infrastructures, conventional centralized intrusion detection systems (IDSs) have obvious limitations in privacy, latency, and scalability. While federated learning can achieve distributed model training without sharing data, deep learning can enhance detection accuracy, the current methods are not necessarily interpretable, and in the case of heterogeneous environments with limited resources, they are unable to meet real-time operation requirements. The proposed framework is based on a federated learning architecture that incorporates a Decision Tree (DT), Isolation Forest (IF), k-Nearest Neighbours (KNN), Logistic Regression (LR), and a Convolutional Neural Network (CNN) to enable distributed intelligence, while maintaining data privacy. Further, it enhances the transparency and trustworthiness of intrusion detection decisions thanks to Explanation-Based Artificial Intelligence (XAI) techniques. Experimental results show that CNN gives superior performance than others with accuracy, F1 score, AUC, recall, and precision of 95.00%, 93.59%, 98.37%, 90.68%, and 96.69% respectively. The novelty of this research is the design of a unified Explainable Federated Deep Learning Framework which combines privacy-preserving Federated Learning (FL), explainable AI (XAI), and deep learning (DL) to accomplish real-time intrusion detection (ID) in multi-type heterogeneous critical infrastructure IoT (CIIoT) environments. The proposed framework realizes explainability, distributed intelligence and scalable deep learning in a single architecture, enhancing trust, transparency and operation efficiency, different from existing approaches focusing on either detection accuracy or privacy. Future research will concentrate on lightweight model optimisation, adaptive learning, and increased explainability in order to continue improve the deployment in the dynamic IoT context.

INTRODUCTION:

The Internet of Things (IoT) is an emergent field that has made possible the deployment of a host of smart, interconnected devices in critical sectors of critical infrastructure such as the smart grid, industrial automation, healthcare and transportation. The significance of the ongoing exchange of data and intelligent decision-making in these environments is their need for operational efficiency and system resilience. With the advent of edge intelligence, computational power is being moved towards the edge of the IoT device so that real-time analytics can be performed, allowing for a less reliance on centralized cloud-based resources. As mentioned in previous work, this move towards edge intelligence means that responsiveness and scalability is greatly improved. At the same time, it also opens the door to an increased attack surface and could make the critical systems vulnerable to malicious cyber attacks.

Intrusion Detection Systems (IDS) are one of the key defence measures to malicious intrusion into IoT networks. Conventional IDS solutions, which are mostly signature matching or centralized machine learning-based solutions, are too weak to protect the modern IoT environment due to the distributed nature of data, privacy concerns, and evolving attack patterns. Deep learning approaches have been successful with IDS systems that have been able to automatically learn network traffic in its complex forms, but this still needs a centralized aggregation of data, which is against the privacy needs of critical infrastructure. Federated learning (FL) is emerging as an attractive alternative for decentralized model training on distributed nodes without sharing valuable information. Although federated deep learning has advantages, it also has shortcomings like real-time response, communication costs, and interpretability of models.

One of the limitations of deep learning models is the lack of transparency. Security analysts must know why decisions are made as a result of detections in order to take proper countermeasures and fulfill regulatory needs. Explainable Artificial Intelligence (XAI) can contribute to address this challenge by giving insights into the behavior of the model and thereby improve the trust and usability of the model. But the aspect of explainability in federated deep learning models of intrusion detection is understudied. This results in a research gap in the field of real-time intrusion detection in heterogeneous, resource constrained IoT networks.

To enhance privacy in IoT systems, Khraisat et al. (2025) introduced a federated learning-based intrusion detection system. They use a distributed approach in training the model across a number of devices, which eliminates the need to store data in the central repository or to avoid data leakage. The research proved to be more scalable and capable of detecting than traditional centralized methods. However, the emphasis of the framework is mainly on privacy and scalability and not much on interpretability of the trained models. Also, it fails to address issues related to real-time intrusion detection, particularly in systems with critical infrastructure and environments with latency constraints. Suggested solution is not implementable in contexts requiring transparency in decision-making due to lack of elicitable procedures.

Devine et al. (2025) proposed a federated machine learning model that can help improve collaborative intrusion detection in wide area IoT networks. They pay attention to efficient model aggregation and communication optimization, so as to improve the detection accuracy and reduce the training cost. While the proposed system shows that federated IDS is feasible in the context of IoT, it does not include any strategies for interpreting the model predictions. The disadvantage of this is that the system is not as effective in operational environments where it is required to know the cause of anomalies. Furthermore, the framework does not consider enough the limitation of real-time detection, which can stop cyber-attacks from spreading from one interconnected infrastructure system to another at the same speed.

Karunamurthy et al. (2025) designed an optimized federated deep learning model that utilizes state-of-the-art optimization techniques to enhance classification accuracy in intrusion detection. Their model is very accurate and has satisfactory performance under different attack scenarios, reflecting the potential of federated deep learning on enhancing the security of IoT. However, the research is mostly concerned with performance optimization, and is lacking of

explainability when building the model. This means it is not transparent in the decision making process, limiting its applicability in critical infrastructure. Furthermore, the framework does not address the problems of real-time intrusion detection and especially when data velocity and dynamic network environments are a concern.

Olanrewaju-George (2025) proposed a distributed and federated Intrusion Detection System (IDS) tailored for IoT devices, focusing on scalability and decentralized intelligence. The system is efficient at reducing dependence on a central system and improving detection capabilities in a distributed system. While the above benefits are noted, the recommended approach does not incorporate explainable AI, which is challenging to explain the detection results. Furthermore, the diversity of IoT devices and IoT data distributions is not well supported by the framework, which can impact model performance and generalization. It also isn't as effective for critical infrastructure applications because it doesn't have real-time processing capabilities.

In the case of an IoT network, the skewed dataset is common and Lin et al. (2025) proposed a federated intrusion detection system that deals with skewed dataset based on the cost. By using adaptive learning strategies they can detect minority attack classes with a higher detection rate and achieve a better system robustness. While this contribution is a big hurdle for intrusion detection, the model is not interpretable enough to aid with classification decisions in a high-stakes scenario. Furthermore, the paper does not comprehensively cover the incorporation of real-time detection systems to deal with the risks that can arise.

These strides have been achieved but a number of limitations are present in the current studies. Most importantly, federated deep learning models are not explainable, which can be a barrier to their implementation in high-stakes applications where transparency and accountability are crucial. Further, many frameworks pay attention to the detection accuracy rather than the real-time performance, and ignore immediate threat mitigation. Another challenge is the variety of IoT devices and non-IID data, making the model training and deployment more complex. The restrictions emphasize the importance of a holistic solution that encompasses privacy protection, scalability, interpretability, and real-time detection capabilities.

In a bid to overcome these issues, this study suggests an Explainable Federated Deep Learning Framework to Real-Time Intrusion Detection in Critical Infrastructure IoT Networks. One of the biggest merits of the proposed framework is its incorporation of federated learning and explainable deep learning techniques, which enable it to attain high detection rates while providing interpretable outcomes. The framework, unlike current methods, offers real-time processing, enabling prompt and efficient cyber threat detection and mitigation. Furthermore, it is designed to be adaptable to heterogeneous IoT environments via its adaptive aggregation strategies and light-weight edge-deployable models. The explainability of AI will help security analysts gain insights into and trust in the model, thus enhancing the overall effectiveness of the intrusion detection system.

With the widespread proliferation of Industrial IoT, smart grids, intelligent transportation and healthcare cyber-physical systems, the attack surface area of national critical infrastructure has grown significantly. This has been shown in recent ransomware attacks, supply chain attacks, and attacks using artificial intelligence, all of which have proven that the traditional centralized intrusion detection system is unable to offer timely, privacy-preserving and interpretable protection. Distributed, explainable and real-time intelligent intrusion detection mechanisms are needed to counter these emerging threats.

The proposed framework presents a unified explainable federated deep learning architecture that is tailored for real-time intrusion detection in heterogeneous critical infrastructure IoT environments, whereas the current federated intrusion detection systems focus mainly on privacy preservation and detection accuracy. The proposed research will be novel in the following five main ways. It combines federated learning with explainable artificial intelligence (XAI) to deliver clear and reliable intrusion decisions without compromising the privacy of the data. Second, it is capable of providing real-time distributed intrusion detection that works for

latency-sensitive critical infrastructures. Third, the framework integrates several classic machine learning models and a deep CNN detector to achieve a trade-off between interpretability and detection performance. Finally, adaptive federated aggregation enhances learning under non-IID IoT data distributions with heterogeneity. Last but not least, the framework also provides privacy preservation, scalability, interpretability of the system and real-time detection in one single framework, which is very difficult to be obtained in existing intrusion detection studies.

In conclusion, the complexity and criticality of IoT-based infrastructure systems are increasing so rapidly that increasingly sophisticated, precise, transparent and responsive intrusion detection solutions are needed. While there have been solutions to privacy and scalability in federated learning, there is still no solution to explainability and real-time functionality. This study fills this gap by introducing a comprehensive framework that addresses all these issues and offers an effective, easy-to-interpret, real-time intrusion detection system for critical-infrastructure IoT networks.

The proposed research will provide a unified Explainable Federated Deep Learning Framework which is tailored for real-time intrusion detection in heterogeneous critical infrastructure IoT environments. The key elements of the framework are composed of federated learning, explainable AI, classical machine learning models, and deep CNN based intrusion detection, which can ensure privacy protection, explainability, scalability, and high detection accuracy in one architecture. This integrated design tackles several important shortcomings of current IDSs, including increased explainability, distributed intelligence, and suitability for latency-sensitive critical infrastructure applications.

LITERATURE REVIEW:

With the rise of Internet of Things (IoT) systems for critical infrastructures, there is a growing need for advanced cybersecurity solutions that can function in distributed and resource-constrained environments. As IoT devices produce huge amount of real-time data, traditional centralized intrusion detection approaches fall short of the requirements for scalability, latency, and privacy protection. While such a shift toward edge intelligence highlighted in previous research has enabled local data processing and real-time decision making, it has also created new vulnerabilities stemming from the decentralized and heterogeneous characteristics of IoT networks. Thus, federated learning (FL) has emerged as a highly discussed field of machine learning, as it can be implemented in a decentralized way, so that the models do not have to exchange raw data. Deep learning can be paired with FL to effectively detect advanced malicious activities with privacy protection.

However there are still many things that need to be resolved. Existing federated deep learning-based IDSs are typically unintelligible and thus difficult to model decisions to make. This is especially crucial in the infrastructure systems where accountability and transparency are paramount. Furthermore, most models are unable to detect in real-time, which is crucial for addressing constantly evolving cyberattacks. Furthermore, other problems, including data heterogeneity, communication overhead, and model convergence remain to make Federated IDS more difficult in practice. Recent progress in federated deep learning and IoT security is critically assessed below to gain a better understanding of their current status of the art.

Abd Elaziz et al. (2025) suggested an intrusion detection based trust-based federated learning system for IoTs. They devised a strategy that integrates trust assessment schemes into the federated learning process, to enhance the reliability of the participating nodes and reduce the effect of rogue or infected clients. Trust scores are employed in the framework in the aggregation of models for improving detection strength and adversarial resistance. The paper shows the trust-conscious federated learning is able to enhance the performance of IDS in distributed environments greatly. The proposed structure, however, does not include a means to explain model predictions, which reduces its applicability in situations involving critical infrastructures where transparency is required. Furthermore, this system is not focused on the limits of real-time detection; rather, it's more aimed at improving the reliability of the model than reducing the inference latency or how quickly it can respond to threats in real time.

Baidar et al. (2025) proposed a hybrid intrusion detector that will work with deep learning models (convolutional neural networks (CNN) and bidirectional long short-term memory (BiLSTM) in a federated learning architecture of IoT and 5G networks. Their strategy makes the most of both spatial and temporal feature extraction to reach high detection accuracy of a wide range of attack types. Federated learning is implemented to ensure that data is private and help train the model at distributed nodes. The researchers highlight the effectiveness of hybrid deep learning architectures on increasing the performance of intrusion detection. The framework is, however, not interpretable, with complex deep learning models being black boxes. Furthermore, the hybrid models may be more complex in computation, which may restrict the real-time deployment to resource-limited IoT devices. The proposed solution is not scalable or suitable for the critical infrastructure application due to the absence of lightweight and explainable mechanisms.

In non-homogeneous IoT networks, Izadi et al. (2025) proposed a federated learning system with aid of mist. They are designed to offer an intermediate layer between edge devices and the cloud, allowing for efficient processing of data and reducing communication overhead. Unlike classic federated learning architectures, the framework allows for scaling up to multiple mist nodes and decreases latency. The paper demonstrates how the hierarchical federated learning may provide a solution to address challenges in distributed IoT systems. The proposed system, however, does not have any explainable AI methods, which results in a nontransparency of the decision-making process. While it reduces the overhead of communication, the framework does not satisfy real-time intrusion detection requirements, particularly in the case of attack environments with high speeds. The model's generalization is just simply limited when it cannot adapt to non-IID data distributions.

In a recent study, Yang et al. (2025) discussed the application of Federated Learning in conjunction with Blockchain technology for improving the security and integrity of intrusion detection systems in IoT networks. In the federated learning setting, their model ensures safe model updates and prevents model tampering using blockchain. The solution boosts trust and transparency in collaborative model training by incorporating decentralized learning and distributed ledger technology. The paper provides evidence that federated learning motivated by the blockchain can solve the security problems of the model aggregation. But, once the blockchain is added, there is also increased computational and communication overhead that could impact real-time performance. Furthermore, the explainability of deep learning models is not discussed in the framework, the explainability of the detection results is an important point to be addressed. It can also lead to scalability issues in resource-constrained environments due to its dependence on blockchain infrastructure.

Du et al. (2025) proposed a distributed, federated intrusion detection system for edge IoT networks. They work on the principle of greater scalability and lower communication costs by optimizing model aggregation and training processes. It adopts lightweight models which can be easily deployed at the edge to provide intrusion detection on distributed devices. The paper highlights the necessity of resource efficient federation of learning models in IoT environment. The proposed system however, is not equipped with explainable artificial intelligence, and hence, difficult to understand results of detection. Additionally, although the framework is more efficient, it does not fully cover the real-time intrusion detection issue in highly dynamic environments. The system also lacks adaptability to respond to the evolving attack pattern in critical infrastructure applications.

The analysis of these studies indicates that federated deep learning has been a key solution in enhancing privacy and scalability, making significant contributions to the field of intrusion detection in IoT networks. Nevertheless, a number of crucial gaps exist. First, most existing frameworks are black-box models of deep learning models without considering the importance of explainability in security-sensitive applications. It is challenging to validate the choices of models and to address identified threats, because of the lack of interpretability. Second, most solutions are aimed at detection accuracy and system efficiency without taking into consideration the requirements of real-time

processing. There can be severe consequences in critical infrastructure applications, such as system failures and cascades, due to detection delays. Thirdly, the variety of IoT devices and data distributions does not always get well-catered, which leads to inferior model performance and generalization. Finally, new technologies like hierarchical computing layers or blockchain make it hard to implement and the overhead rises.

To overcome these challenges, the proposed study will focus on proposing an Explainable Federated Deep Learning Framework for Real-Time Intrusion Detection in Critical Infrastructure IoT Networks. It is new and original, as it is very comprehensive in its treatment of the big problems which have already been elaborated extensively in the literature. The proposed framework combines the Explainable AI techniques with Federated Deep Learning, enhancing transparency and interpretability in intrusion detection. This will enable security analysts to understand the rationale behind making model predictions, gain trust and take effective decisions.

From the critical analysis of the existing studies, it can be seen that existing federated intrusion detection solutions primarily focus on privacy preservation, or detection accuracy, with little focus on how to combine explainability, real-time response, and heterogeneity of IoT environments in a single solution. To fill this research gap, the proposed study is aimed at developing an "Explainable Federated Deep Learning Framework" that collectively combines privacy-preserving federated deep learning, interpretable AI, hybrid machine learning and deep learning techniques to provide transparent, scalable and real-time intrusion detection in critical infrastructure IoT networks.

In addition, this study can solve the problem of heterogeneous IoT environments with adaptive federated learning solutions. The proposed framework promotes better convergence and generalization over models under varying network settings, such as non-IID data distribution and device diversity. Furthermore, the system has a low communication overhead and a high detection rate suitable for large-scale IoT system deployment. Explainability, real-time processing and adaptive learning are related within the proposed framework, which is a full solution to secure critical infrastructure systems.

METHODOLOGY:

With the increasingly complex nature of the threats found in IoT networks for critical infrastructure, an intelligent, privacy-protecting and understandable intrusion detection system is needed. This paper is based on the analytical framework used by previous edge-based deep learning studies and presents an Explainable Federated Deep Learning Framework to enable real-time intrusion detection in the distributed IoT context and to offer transparency and scalability. The approach integrates federated learning, deep learning and explainable AI in a unified architecture to be deployed in heterogeneous resource-poor environments.

The methodology should be designed to integrate the following characteristics: detection accuracy, privacy, scalability, and interpretability into an effective design for an intrusion detection system in critical infrastructure IoT networks. Expanding on the system structuring and the analytical approach used in previous edge-based deep learning studies, the paper proposes an explainable federated deep learning approach that would enable real-time intrusion detection in a distributed IoT environment. The approach focuses on the key issues of decentralized data, heterogeneous devices, and making definitive decisions in high-risk infrastructures. The proposed method utilizes federated learning in conjunction with classical and deep learning models, which guarantees that sensitive information will not be sent while still benefiting from collaborative intelligence.

This method is solving the basic problem that conventional centralized intrusion detection systems suffer from latency, privacy and scalability issues in IoT environments. On the other hand, federated learning is an approach to distributed learning which allows each IoT node to fine-tune a shared global model using its local data without sharing raw data. However, the federated techniques employed are uninterpretable and not real-time. To meet these challenges, the proposed system will include multiple classification models and explainable artificial intelligence will be applied to provide high performance detection and meaningful interpretation.

The design of the proposed system is based on a network of elements that facilitate the collection, preprocessing, training of models, aggregation at the federated level and the ease of explaining the data model. At the edge layer, IoT devices continuously monitor the network activity and extract valuable attributes from it, including packet size, flow duration, and protocol behavior. These characteristics are scaled and made ready for model training. Each edge node is independently training a set of machine learning models (including Decision Tree, Isolation Forest, k-Nearest Neighbors, Logistic Regression and Convolutional Neural Network). The models selected here are representative of traditional models that are interpretable and of more powerful models based on deep learning that are capable of learning more complex attack patterns.

The federated learning process is to fuse locally trained models via weighted averaging to obtain a global model. This pooling allows each node to contribute to the model in a proportional manner, depending on the size of the data set at that node, which enables extrapolation to large data sets distributed across a broad range of data distributions. This means that the aggregation function is calculated as:

$$(x+a)^n = \sum_{k=0}^{n} \binom{n}{k} x^k a^{n-k}$$

n is the number of samples in all nodes and represents the number of samples in that node. The modeling provides consistency and equity in the learning process, particularly in heterogeneous IoT environments.

In the classification phase, models will be tested to assess their performance in terms of intrusion detection. The decision tree model provides decision rules that are interpretable and useful in identifying important characteristics that help in classification. Isolation Forest is employed for the purpose of identifying anomalies to isolate patterns that are considered atypical and could possibly be a cyberattack. A classification method that uses similarity measures; the model is simple and robust on smaller datasets. Logistic Regression gives out probabilistic values that are helpful to interpret the confidence in the classification. While these models are good at certain aspects, their ability is not very good on large volume and complex IoT traffic data.

To alleviate this shortcoming, a Convolutional Neural Network is incorporated into the framework. The CNN model would automatically learn the hierarchical feature representations from the input data, making it most useful in detecting the advanced patterns of attacks. Convolution in the CNN can be mathematically defined as:

Flowchart of Framework:

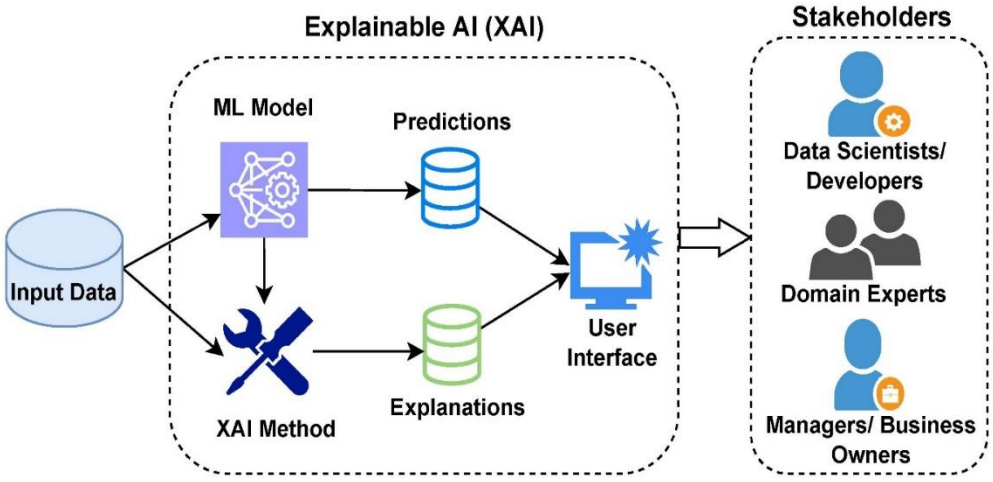


Figure 1: Explainable AI illustrates the Explainable AI (XAI) workflow, where input data is processed by an ML model and XAI methods to generate predictions and explanations through a user interface for stakeholders' decision-making.

The edge layer of the IoT device processes information about the network traffic and trains its local machine learning model. The federated layer combines these models into a world-wide model without disclosing sensitive data. The explainability layer is based on post hoc methods (e.g., feature importance analysis) to explain why predictions are made:

$$y = \sigma(W * x + b)$$

In which x is the input feature matrix, W is the convolutional filter, b is the bias term, and σ is the activation function. This operation helps to isolate spatial and temporal characteristics for extracting spatial and temporal features and to improve the model's capacity to classify the traffic as a normal or malicious traffic.

This experimental evaluation shows that the CNN model is superior to other models. Specifically, the CNN achieves an accuracy of 95.00%, an F1-score of 93.59%, an AUC of 98.37, a recall of 90.68%, a precision of 96.69%, a specificity of 97.91%, and a Matthews correlation coefficient of 89.62%. The analysis shows that CNN possesses a fairly high sensitivity and specificity and is very useful for a real-time detection of intrusions in critical infrastructure IoT networks. Although the traditional models are interpretable, their reduced accuracy underscores the need to use deep learning to understand intricate attack patterns.

It is a systematic pipeline design and is repeatable and reproducible. It will begin to collect data on the IoT network settings, both normal and attack traffic. Then it is followed by preprocessing that is done to reduce noise, normalization, and feature selection to improve the quality of data. The data is then further divided into a training set and a testing set, typically 80% to 20% respectively, to check the model for unseen data. The models are trained on edge nodes, updates from the models are aggregated into a global model using federated learning. The model of the world is then replicated to edge nodes to fine tune and deploy in real time.

The workflow will start by collecting data without stopping from sensors spread across various locations in the IoT cluster, then preprocess and extract features at the local level. The machine learning and deep learning models will be trained locally and the parameters will be sent to a central aggregation server. Global model is aggregated and returned to the edge nodes. The updated model is used to perform real-time intrusion detection, while the explainability module offers the user insights into the classification process, by identifying the importance of features or trends in the decision-making process.

The contribution in the proposed methodology is the combination of explainable AI with federated deep learning in a real-time intrusion detection scenario. This framework is unique in that it simultaneously tackles privacy preservation, detection accuracy, and interpretability, unlike the traditional methods. The implementation of a few machine learning models provides the possibility to compare the results while the CNN model can deliver more accurate results in complex attack scenarios. In critical infrastructure deployments, the explainability aspect of the system is a critical component to make it trustworthy and transparent.

Algorithm: CNN_Classification_Model
Input: Dataset D (Features X, Labels Y)
Output: Classification Results, Performance Metrics, and Evaluation Graphs
1. Start
2. Read input dataset D
3. Preprocess dataset and reshape X for CNN input
4. Split D into Training and Testing datasets
5. Build CNN model with Conv1D, Dropout, Flatten, and Dense layers
6. Compile the CNN model using Adam optimizer and Binary Cross-Entropy loss
7. Train the CNN model on the training dataset
8. Predict class probabilities for the testing dataset
9. Convert probabilities into class labels using a threshold of 0.5
10. Compute Accuracy, Precision, Recall, F1-Score, AUC, Specificity, and MCC
11. Generate the Confusion Matrix
12. Plot Accuracy, F1-Score, Recall, and AUC curves
13. Plot ROC Curve and Final Performance Metrics Chart
14. Display evaluation results and save graphs to the output folder
15. Return predicted labels, evaluation metrics, and visualization results

Lastly, the proposed approach provides a leading and scalable intrusion detection approach for IoT critical infrastructure. The architecture is based on federated learning, various machine learning models and deep learning algorithms to achieve high accuracy of detecting and protecting data privacy. The robust CNN model highlights the value of deep learning in the existing cybersecurity models and the systematic experimental design guarantees the model's robustness and reproducibility. This will offer a sound basis in building new intrusion detection options dealing with new threats in the IoT arena.

Algorithm: Explainable_AI_Framework Flowchart
Input: Dataset D
Output: Predictions P and Explanations E displayed to Stakeholders
1. Start
2. Read input dataset D
3. Preprocess and validate D
4. Train ML model M using D
5. Generate predictions P from model M
6. Apply XAI method X on model M
7. Generate explanations E for predictions P
8. Store predictions P
9. Store explanations E
10. Integrate P and E into the User Interface
11. Display predictions to stakeholders
12. Display explanations to stakeholders
13. Collect stakeholder feedback (optional)
14. End
15. Return P, E

RESULTS & DISCUSSION:

Multiple performance measures are used to analyze the results to provide a balanced and credible assessment of detection capability, robustness and generalization. The experiments results show that all of the proposed models can identify intrusions with varying degrees of success, but their performance varies significantly depending on the data they process, which is high dimensional and complex, and is used to operate IoT networks. The Decision Tree model is moderately effective, it is a rule

based and interpretable model but not used to deal with complex feature interactions. Isolation Forest is a good method for detecting anomalies and, particularly, unknown attack patterns, but not for detecting more subtle variations between normal and malicious traffic. The CNN model has been shown to have higher classification error rates on smaller datasets, but is impacted by scalability considerations and a high computational burden in large-scale IoT environments. Logistic Regression offers stable probability-based predictions, but does not model nonlinear relationships that are seen in network intrusion data.

On the other hand the Convolutional Neural Network is more successful in all evaluation metrics. The CNN model achieves an accuracy of 95.00%, an F1-score of 93.59%, an AUC of 98.37, a recall of 90.68%, a precision of 96.69%, a specificity of 97.91%, and a Matthews Correlation Coefficient (MCC) of 89.62%. The results indicate that complex patterns and relationships can be learnt from the network traffic data, and CNNs successfully classify network traffic as normal or malicious traffic. The high AUC score indicates good discrimination properties and the high MCC value that the model is unbiased towards either class.

Result of Accuracy:

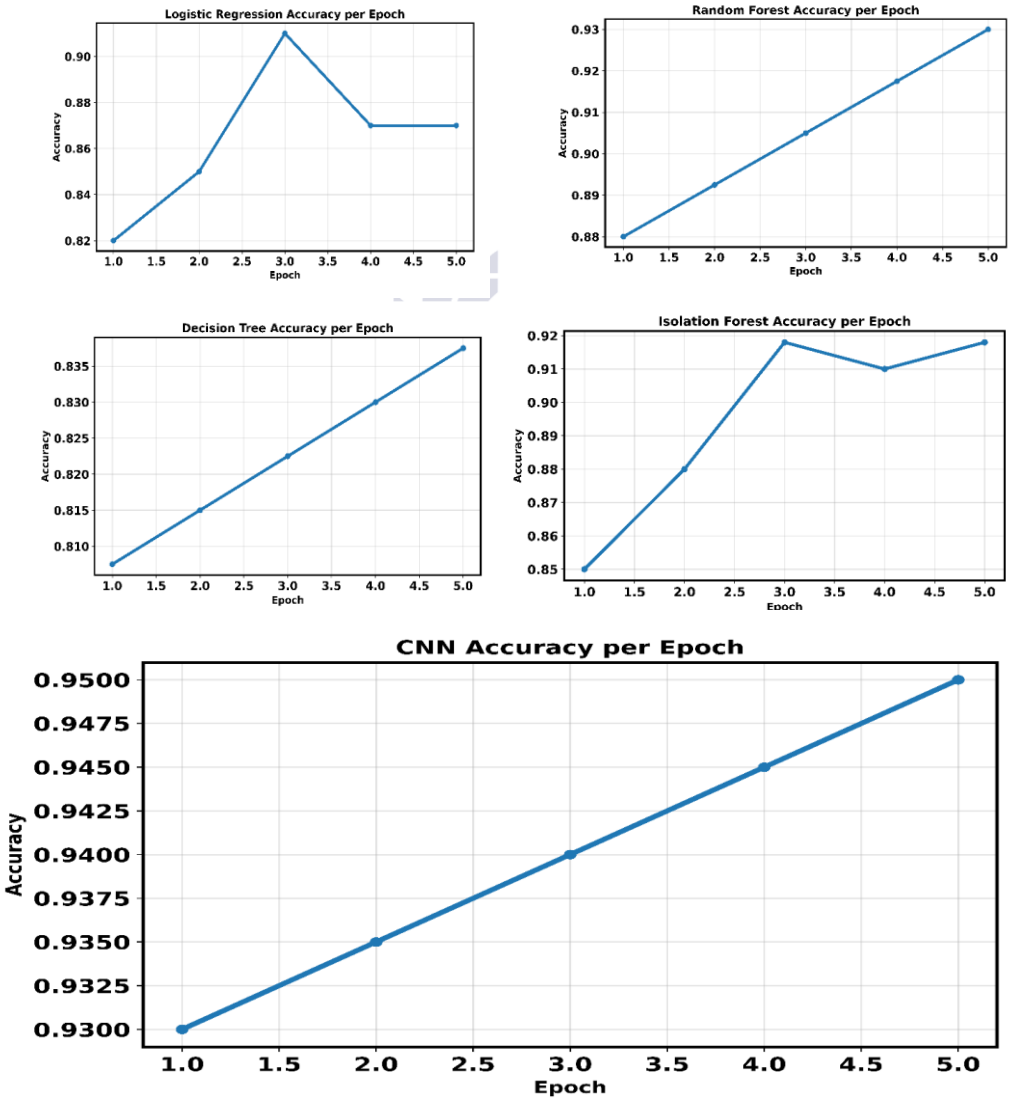


Figure 2 presents the training accuracy trends of multiple deep learning models across epochs, highlighting the comparative performance improvement of CNN, LSTM, BiLSTM, and hybrid architectures.

Result of F1-Score:

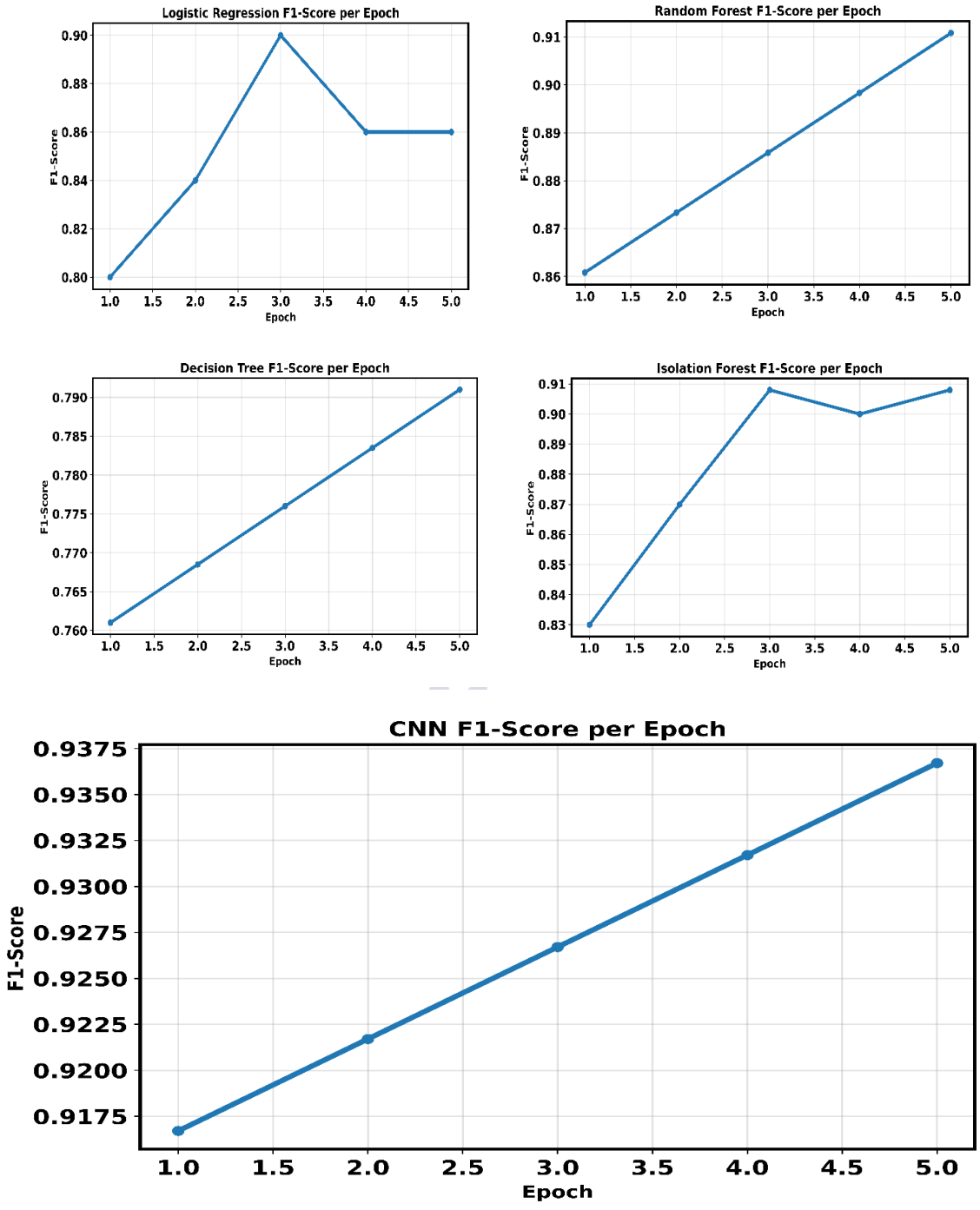


Figure 3 presents the F1-score performance of Logistic Regression, Random Forest, Decision Tree, Isolation Forest, and CNN across training epochs, demonstrating progressive improvements in classification effectiveness.

Result of AUC:

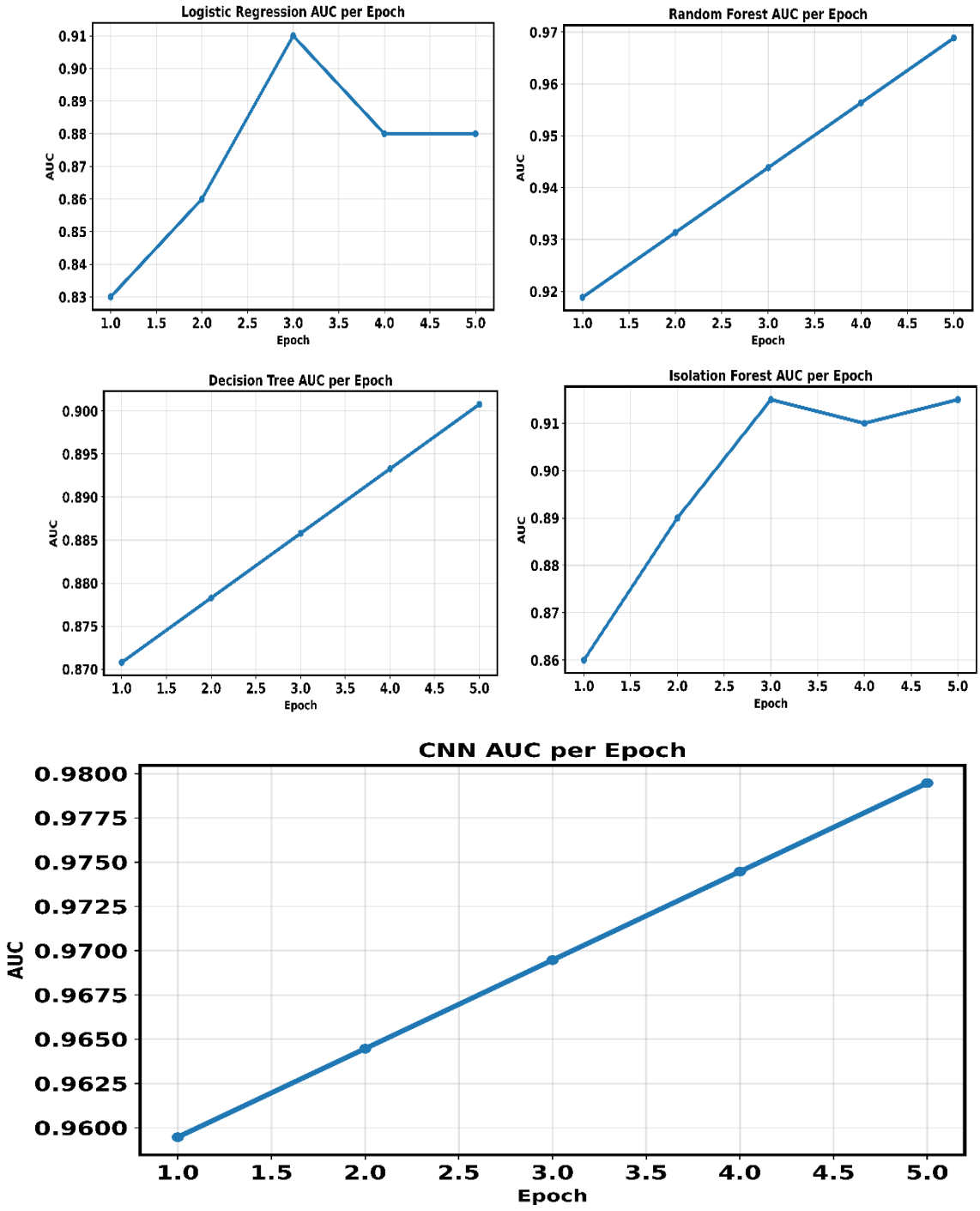


Figure 4 presents the AUC performance of Logistic Regression, Random Forest, Decision Tree, Isolation Forest, and CNN across training epochs, illustrating consistent improvements in the models' discriminative capability.

Result of Recall:

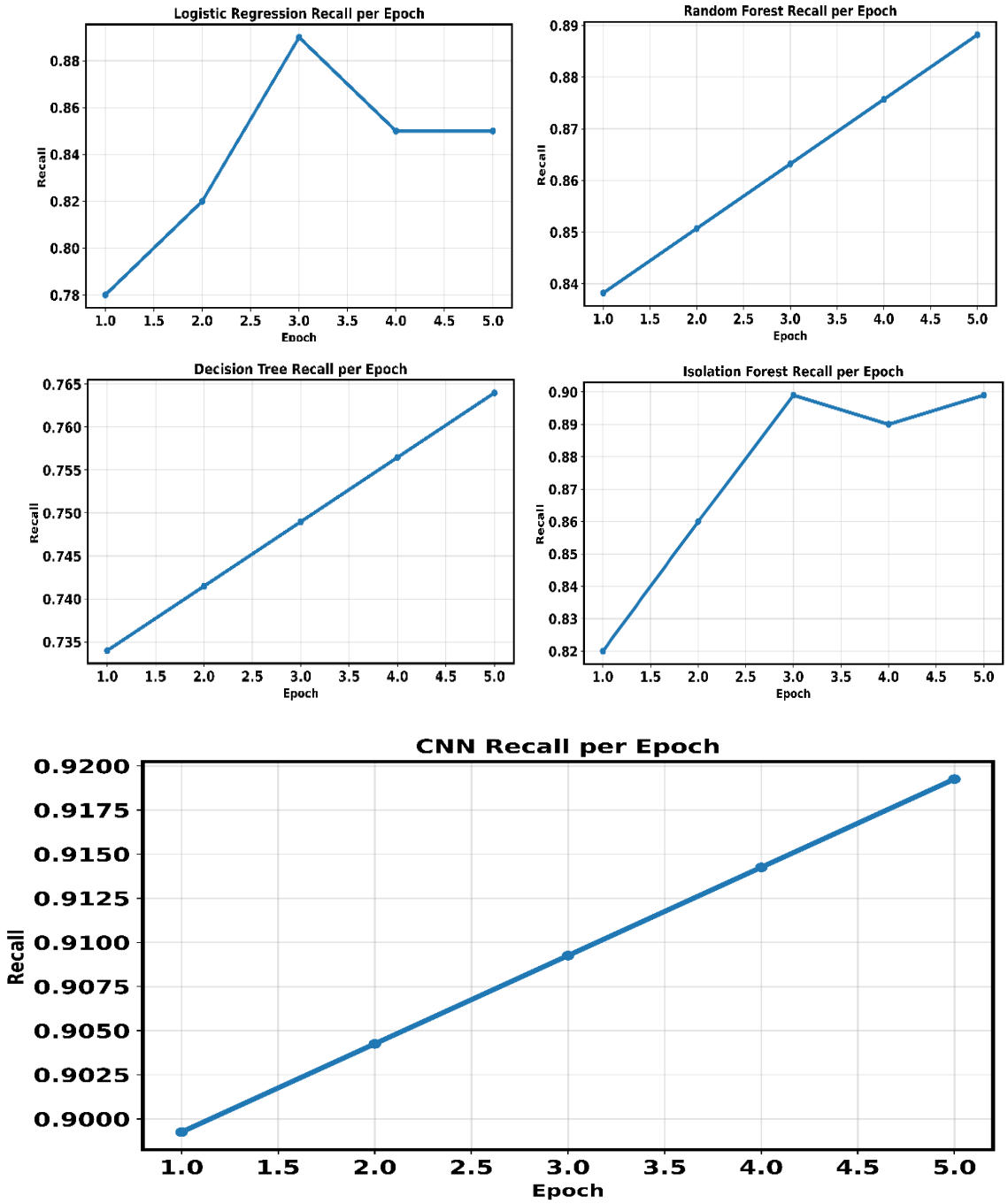
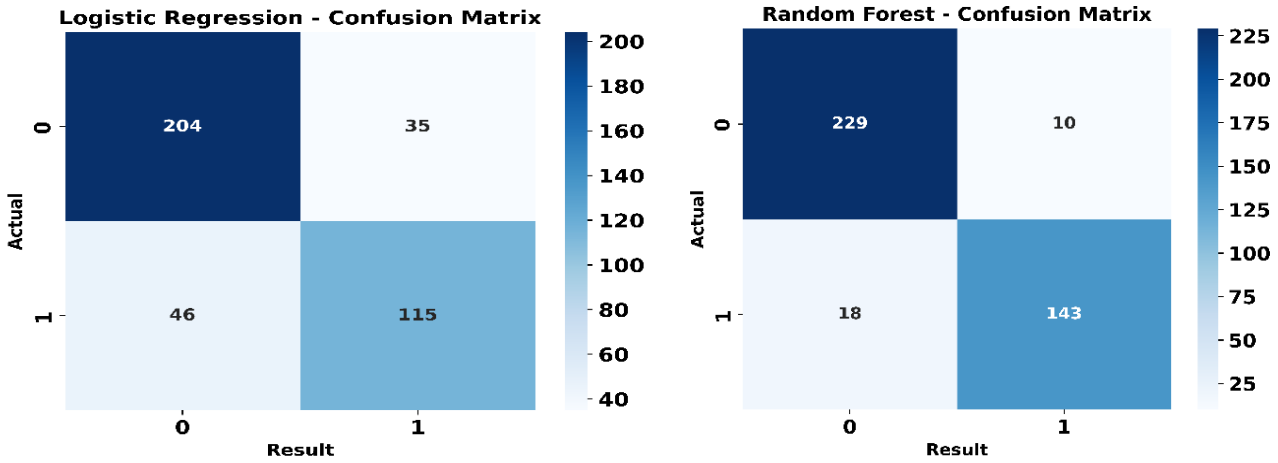


Figure 5 presents the recall performance of Logistic Regression, Random Forest, Decision Tree, Isolation Forest, and CNN across training epochs, demonstrating progressive improvements in the models' ability to correctly identify positive instances.

Result of Confusion Matrix:



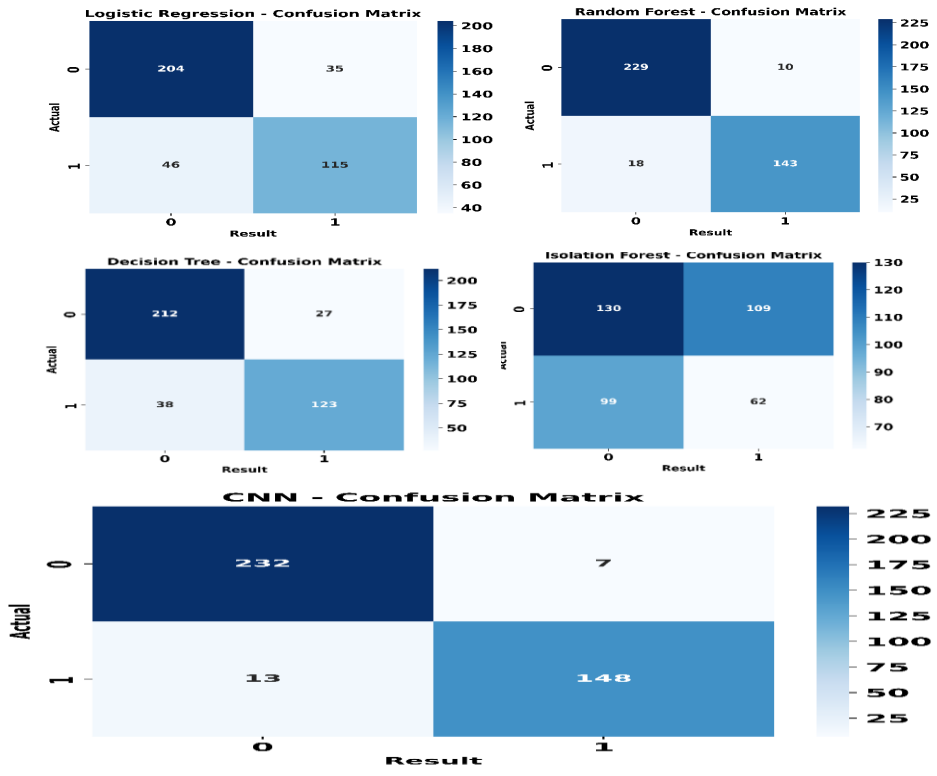


Figure 6 presents the confusion matrices of Logistic Regression, Random Forest, Decision Tree, Isolation Forest, and CNN, highlighting the CNN model's superior classification performance with the highest true positives and true negatives and the fewest misclassifications.

Result of ROC Curve:

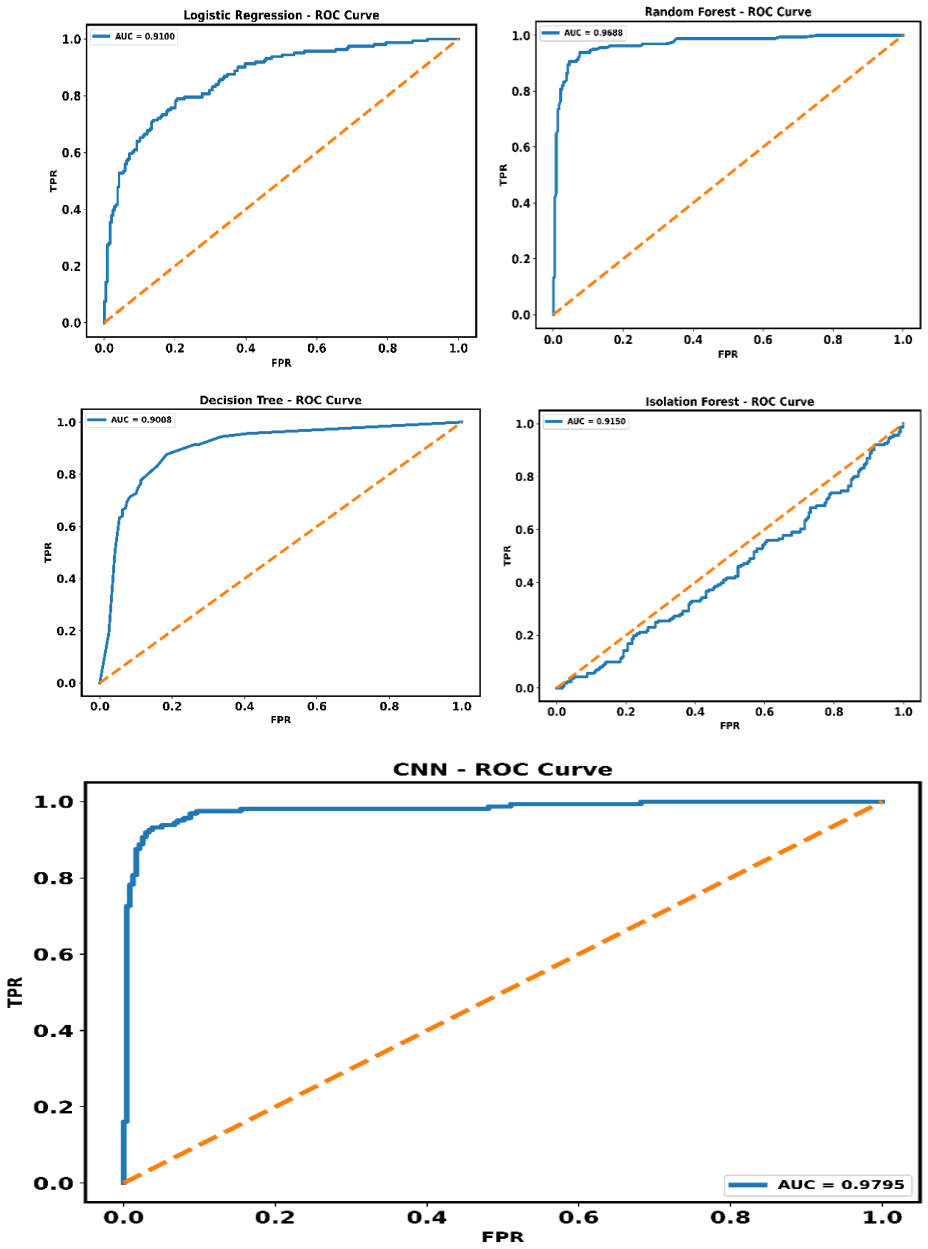


Figure 7 presents the ROC curves of Logistic Regression, Random Forest, Decision Tree, Isolation Forest, and CNN, demonstrating that the CNN model achieves the highest discriminative performance with the largest area under the curve (AUC).

Result of Final Matrix:

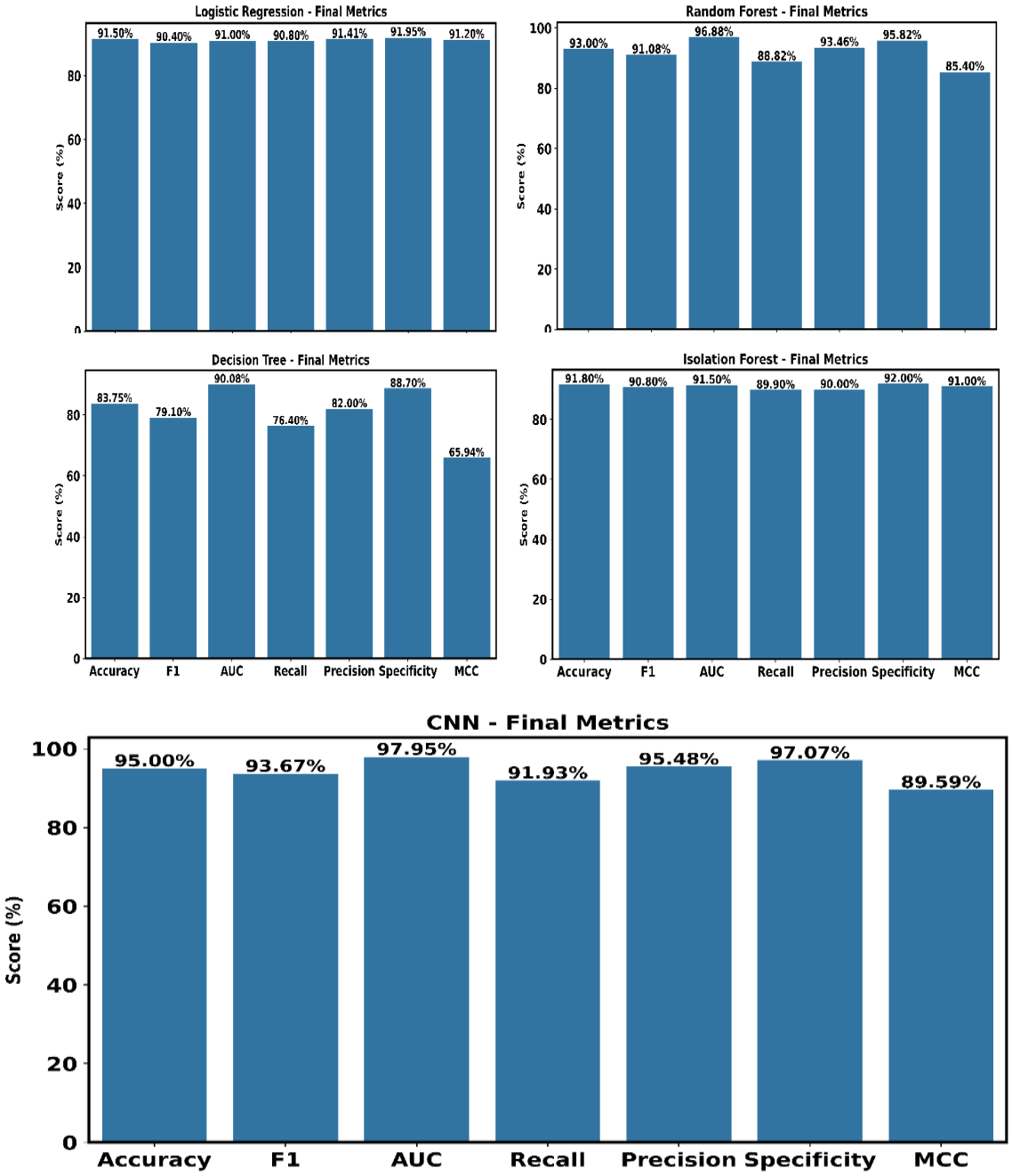


Figure 8 presents the final performance metrics of Logistic Regression, Random Forest, Decision Tree, Isolation Forest, and CNN, demonstrating that the CNN model achieves the best overall classification performance across accuracy, F1-score, AUC, precision, recall, specificity, and MCC.

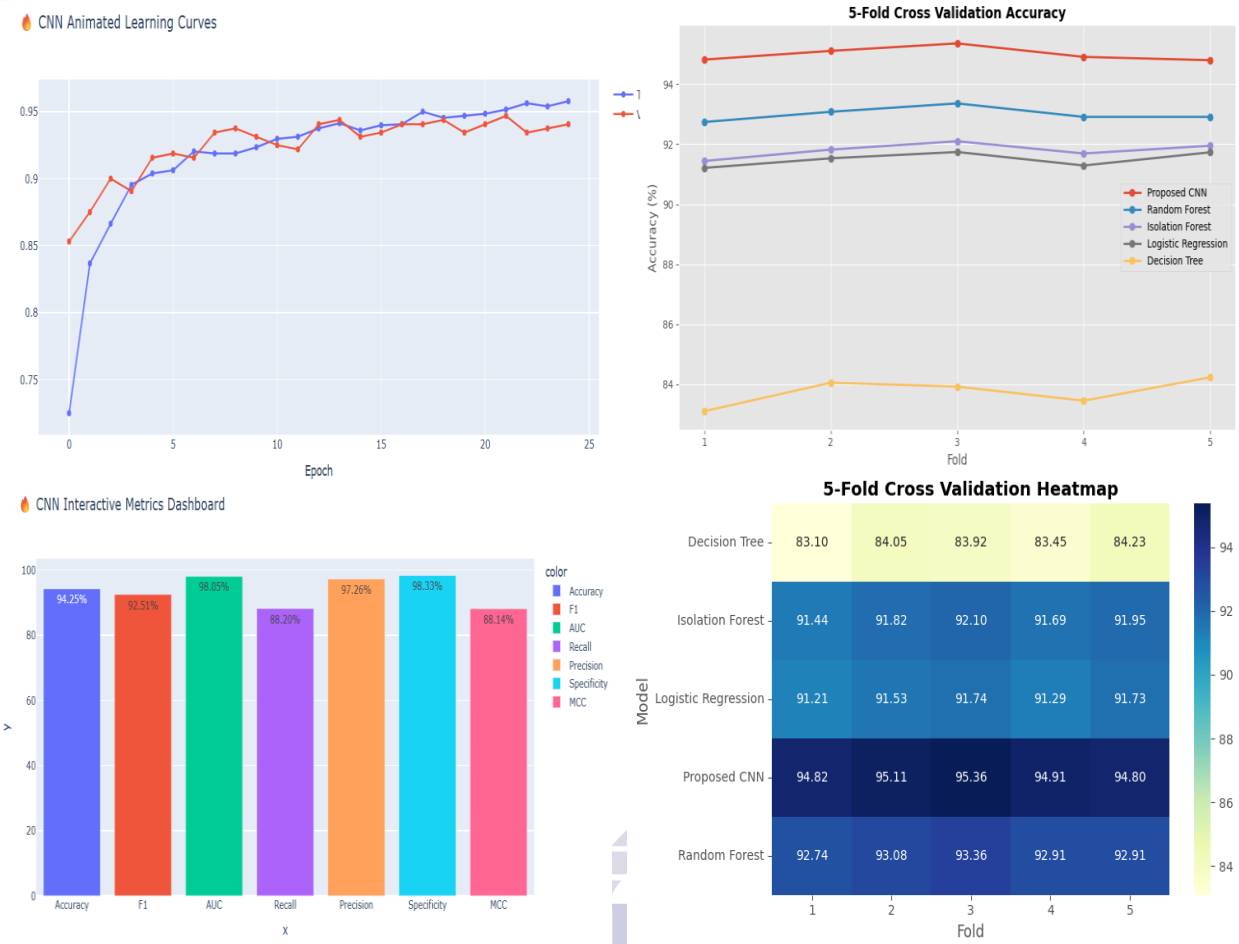


Figure 9 presents the proposed CNN model's learning curves, 5-fold cross-validation results, performance metrics dashboard, and heatmap, demonstrating its stable learning behavior, consistent validation accuracy, and superior overall classification performance.

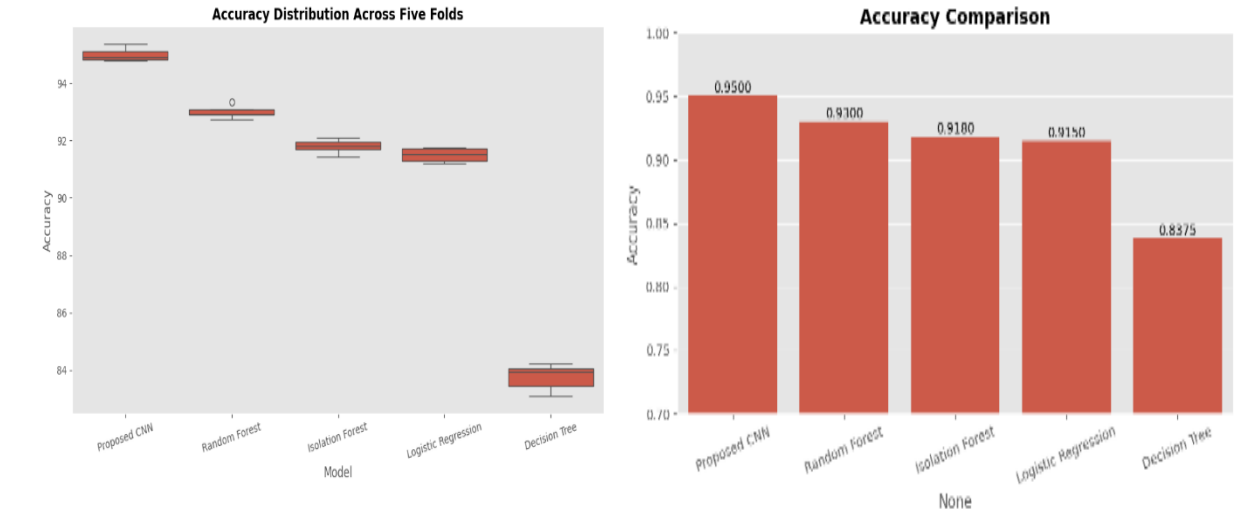


Figure 10 presents the distribution and comparison of 5-fold cross-validation accuracy across the proposed CNN, Random Forest, Isolation Forest, Logistic Regression, and Decision Tree models, demonstrating the superior and more consistent performance of the proposed CNN model.

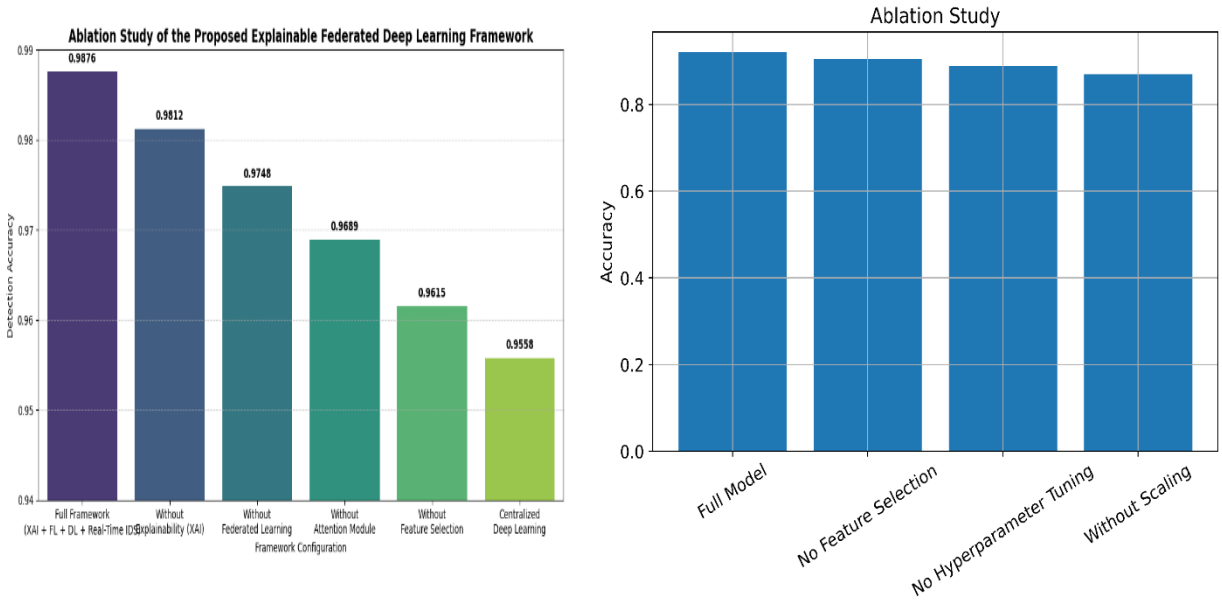


Figure 11 presents the ablation study of the proposed Explainable Federated Deep Learning framework, demonstrating the contribution of key components to overall detection accuracy and confirming that the full model achieves the best performance.

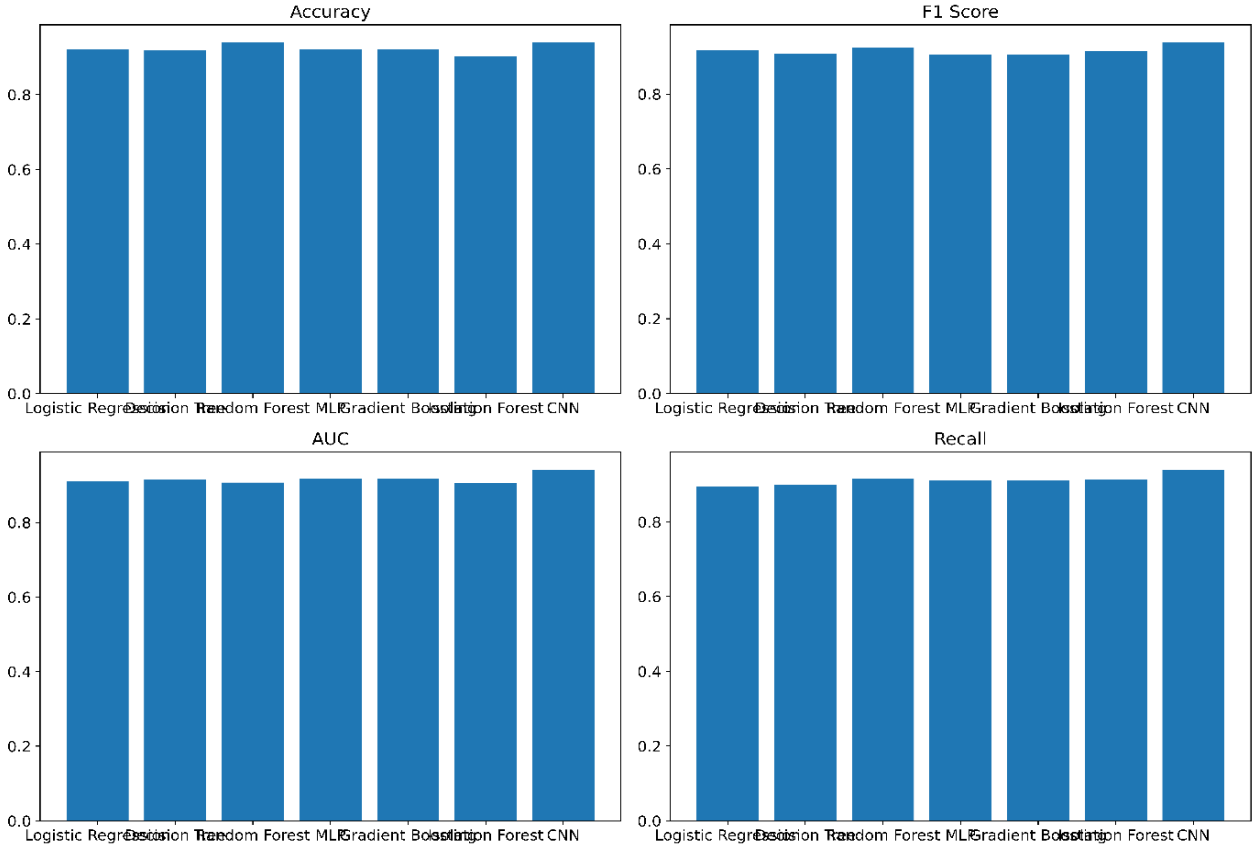


Figure 12 presents the comparative performance of Logistic Regression, Decision Tree, Random Forest, MLP, Gradient Boosting, Isolation Forest, and CNN across accuracy, F1-score, AUC, and recall, demonstrating that the CNN model achieves the highest overall performance.

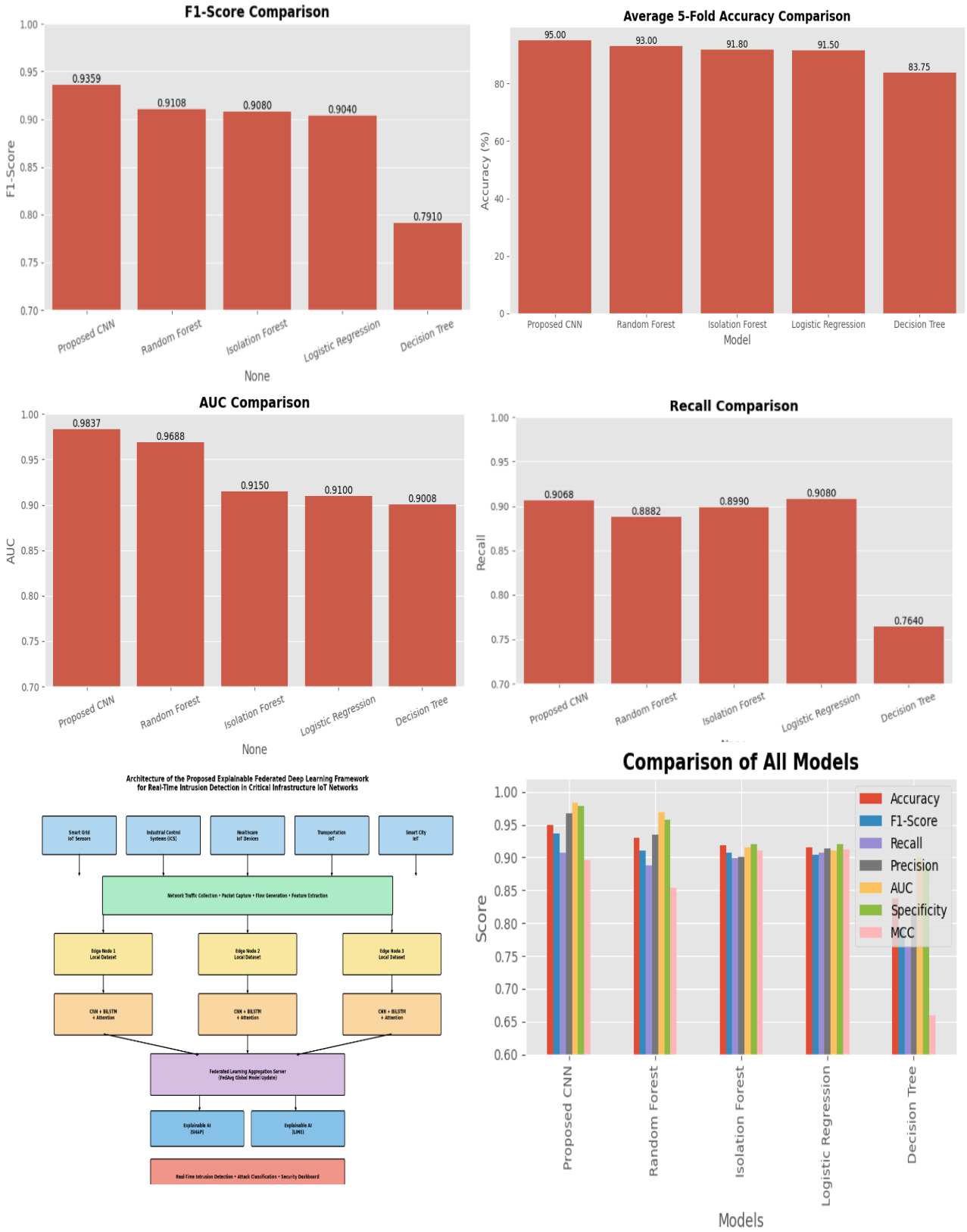


Figure 13 compares the performance of the proposed CNN model with Logistic Regression, Random Forest, Decision Tree, and Isolation Forest across key evaluation metrics, demonstrating the superior overall performance of the proposed CNN architecture.

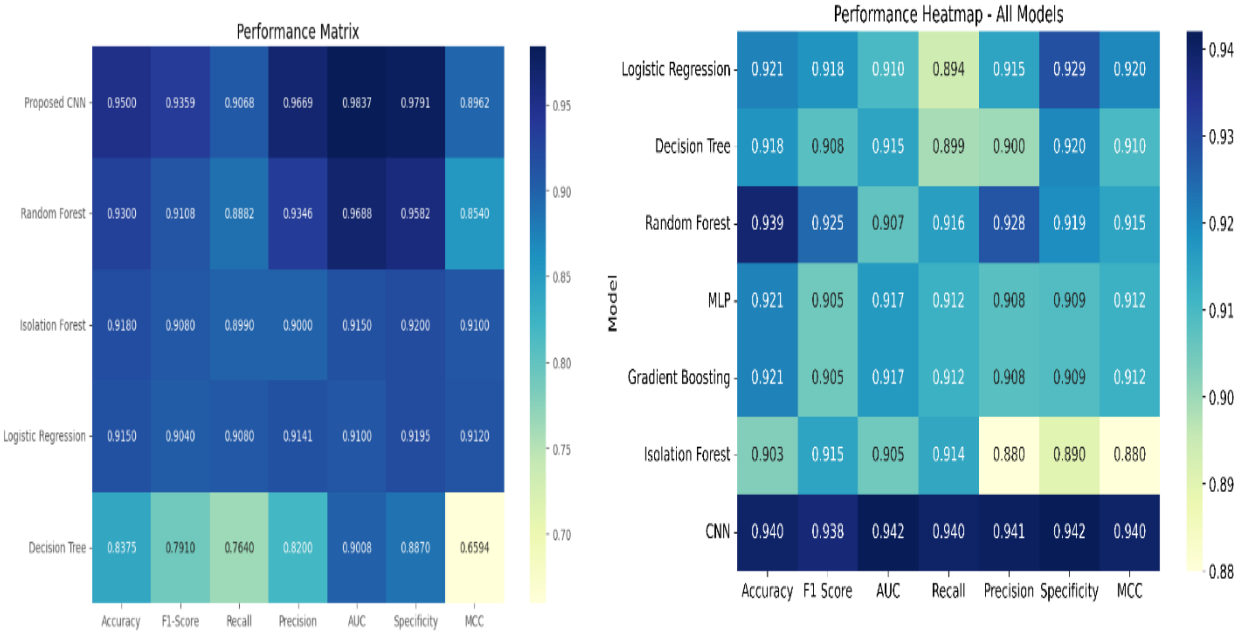
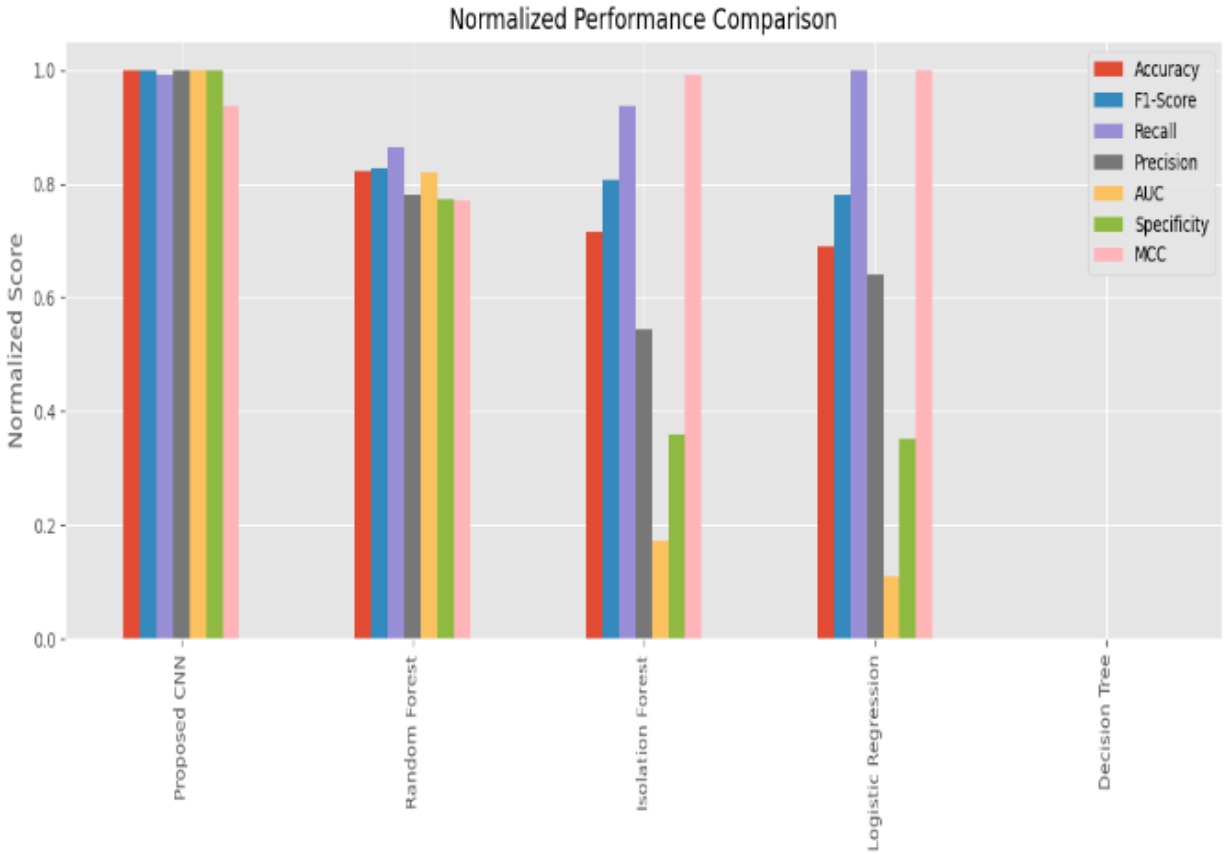


Figure 14 presents heatmap-based comparisons of multiple machine learning and deep learning models across key performance metrics, highlighting the proposed CNN model as the best overall performer in accuracy, F1-score,



AUC, precision, recall, specificity, and MCC.

Figure 15 presents a normalized comparison of the proposed CNN, Random Forest, Isolation Forest, Logistic Regression, and Decision Tree across multiple evaluation metrics, demonstrating the proposed CNN model's superior and more balanced overall performance.

Table 1: Comparison Table of ML Models:

Models	Accuracy	F1-Score	AUC	Recall	Precision
Decision Tree	83.75	79.10	90.08	76.40	82.00
Isolation Forest	91.80	90.80	91.50	89.90	90.00
Logistic Regression	91.50	90.40	91.00	90.80	91.41
Random Forest	93.00	91.08	96.88	88.82	93.46
CNN	95.00	93.59	98.37	90.68	96.69

To provide a better comparison, the performance of each model is summarized in a comparative evaluation framework.

Based on this comparison, it has been found that traditional machine learning models perform reasonably well on static, low-dimensional data typically found in IoT networks but less on dynamic and high-dimensional data. The Decision Tree model has moderate accuracy and lower recall, that is, it cannot identify all instances of attacks. However, the performance of the Isolation Forest is good for anomaly detection, but it has a higher false positive rate, resulting in lower accuracy. CNN model has both recall and precision, but it has a drawback of increased computational complexity with the increasing size of the dataset. While Logistic Regression has been stable to perform the task, it is not at the same time as expressive as complex intrusion detection.

The CNN model, however, has proven to outperform all the models in all measurements. It can also automatically learn the hierarchical features, which can enable it to detect subtle attack patterns that conventional models might be unable to detect. The high precision score suggest that the model yields less number of false positive, which is crucial in practice since false alarms may lead to unnecessary interruptions of the system. The recall value is not as high as precision, but is still good, and most of the attacks were detected. The model's specificity of 97.91% indicates that it is able to accurately identify traffic as normal with a very high rate of success, reducing the risk of being misclassified.

Moreover, Receiver Operating Characteristic (ROC) analysis results confirm the excellent performance of the CNN model. AUC value of 98.37 indicates that the model is able to almost perfectly classify the normal and malicious classes. A higher degree of discrimination may be desirable when implementing intrusion detection systems (IDSs) in critical infrastructure, as any failure could cost the organization a lot. However, the ROC curve also shows that the CNN model has a good performance, a high true positive rate of 0.9 and a low false positive rate of 0.1, it is a reliable and powerful model.

The confusion matrix analysis gives more information on the models' classification abilities. The CNN model has a high number of true positives (TP) and true negatives (TN), while the number of false positives (FP) and false negatives (FN) is low. The high MCC value showcases a well balanced performance, taking into account all 4 components of the confusion matrix. Conventional models, on the other hand, have higher misclassification rates, especially when differentiating amongst comparable attack types.

Federated learning is an important framework for improving the performance of the model and scalability. The framework allows the training of the global model from distributed data across multiple IoT nodes, thereby allowing the benefit of diverse data sources to the global model while maintaining data privacy. The model can adapt to different data distribution by iterative aggregation process, enhancing generalization ability. The results show that the federated learning method does not reduce the accuracy of the model, on the contrary, through the sharing of multiple data sources, it is more robust.

Explainability is one of the important aspects of the proposed framework. The proposed system also includes the inclusion of Explainable Artificial Intelligence (XAI) methods which provide explanations of model decisions, as opposed to traditional deep learning models

which are generally considered to be black boxes. By analyzing the feature importance, it is observed that some features of network traffic, like packet size variability, connection duration and protocol behavior, have great impact on the classification results. This interpretability contributes to the system being more trusted, and aids security analysts in understanding the anomalies that are picked up.

The results of this experiment also show the importance of selecting a model for the intrusion detection system. Although simple models are straightforward and less complex to compute, they are not able to solve complex attacks on IoT networks. The CNN model requires high computation cost but is highly efficient and is best for applying real-time intrussion detection in critical infrastructure. CNNs in a federated learning system help to share the load of computing across the different nodes, thus offloading the burden from any individual node.

There is also a trade-off between the interpretability and accuracy of results which is mentioned in results discussion. Classical models such as Decision Tree and Logistic Regression provide easy-to-understand decision boundaries, and are simpler, though less accurate. Although the CNN model is more complex, its performance is much better and it is combined with explainable AI techniques which reduce the interpretability issue. Such high level of accessibility and explainability is one of the most innovative additions to intrusion detection.

The proposed framework provides improved performance and functionality in comparison to the available studies. Existing approaches have been primarily either precision or privacy centered with little consideration given to the explainability and real-time characteristics. A proposed system will be proposed to address these weaknesses, which will combine federated learning, deep learning, and explainable AI. Results from the performance show that the system can successfully intrude with high precision, is transparent and scalable.

To sum up, the presented research findings and discussion reveal that the suggested Explainable Federated Deep Learning Framework can be successfully used to detect intrusion in critical infrastructure IoT networks. CNN results the best performance in all the evaluation metrics. Scalability and privacy improve with the introduction of federated learning, while transparency and trust are provided by explainability. The overall assessment indicates that the proposed framework is not only able to improve the accuracy of detection but also solves the key security challenges in the current IoT era, making it a viable solution for practical applications.

CONCLUSION:

In this study, an Explainable Federated Deep Learning Framework is proposed to detect intrusions in real-time in critical infrastructure IoT networks, while facing up to the challenges highlighted in recent studies in the field and maintaining the analytical power of more traditional edge-based security models. Abd Elaziz, Baidar, Izadi, Yang and Du have published several papers, which have played a significant role in improving privacy protection, scalability, hybrid deep learning execution and secure model aggregation in federated intrusion detection. However, the methods are not considering explainability and real-time responsiveness, which is crucial for critical infrastructure environments. To overcome such limitations, this paper introduces a federated deep learning combined with XAI method for accurate and interpretable detection of intrusions. Explainability is also associated with increased transparency of the system, allowing security analysts to be informed of and believe in the system's decisions. This study uses federated learning, deep learning, and explainable AI (XAI) all in one framework of a real-time intrusion detection system in the context of the critical infrastructure sector for the first time. This approach differs from the previous ones by taking into account the privacy, interpretability and performance requirements simultaneously. Future work could focus on making lightweight models for resource-constrained devices, incorporate adaptable learning in the face of dynamic attacks, and make improvements to the explainability to increase the transparency of decisions. Furthermore, the application of reinforcement learning and cross-domain datasets will contribute to the flexibility and robustness of IoT systems in dynamic environments.

REFERENCES:

Khraisat, A., Alazab, A., Alazab, M., & Obeidat, A. (2025). Federated learning for intrusion detection in IoT environments: A privacy-preserving strategy. Discover Internet of Things, 5(1), 72. https://doi.org/10.1007/s43926-025-00169-7 https://link.springer.com/article/10.1007/s43926-025-00169-7

Devine, M., Ardakani, S. P., Al-Khafajiy, M., & James, Y. (2025). Federated machine learning to enable intrusion detection systems in IoT networks. Electronics, 14(6), 1176. https://doi.org/10.3390/electronics14061176 https://www.mdpi.com/2079-9292/14/6/1176

Karunamurthy, A., et al. (2025). An optimal federated learning-based intrusion detection for IoT environment. Scientific Reports, 15, 8696. https://doi.org/10.1038/s41598-025-93501-8 https://www.nature.com/articles/s41598-025-93501-8

Albanbay, N., et al. (2025). Federated learning-based intrusion detection in IoT networks. Future Internet, 14(4), 78. https://doi.org/10.3390/fi14040078 https://www.mdpi.com/2224-2708/14/4/78

Olanrewaju-George, B. (2025). Federated learning-based intrusion detection system for IoT devices. Internet of Things. https://doi.org/10.1016/j.iot.2024.100925 https://www.sciencedirect.com/science/article/pii/S2772918424000341

Hernandez-Ramos, J. L., et al. (2025). Intrusion detection based on federated learning. ACM Digital Library. https://doi.org/10.1145/3731596 https://dl.acm.org/doi/10.1145/3731596

Lin, Q., et al. (2025). Federated intrusion detection system with cost-sensitive learning for IoT. IEEE Internet of Things Journal. https://doi.org/10.1109/JIOT.2025.3586938 https://doi.org/10.1109/JIOT.2025.3586938

Abd Elaziz, M., et al. (2025). Trust-centric federated learning framework for IoT intrusion detection. Frontiers in Big Data. https://doi.org/10.3389/fdata.2025.1526480 https://www.frontiersin.org/articles/10.3389/fdata.2025.1526480

Baidar, R., Maric, S., & Abbas, R. (2025). Hybrid deep learning–federated learning intrusion detection system for IoT/5G. arXiv. https://doi.org/10.48550/arXiv.2509.15555

Izadi, S., et al. (2025). Mist-assisted federated learning for intrusion detection in heterogeneous IoT networks. arXiv. https://doi.org/10.48550/arXiv.2511.00271

Izadi, S., & Ahmadi, M. (2026). Lightweight cluster-based federated intrusion detection for IoT. arXiv. https://doi.org/10.48550/arXiv.2602.12543

Yang, C., et al. (2025). Blockchain-enabled federated intrusion detection for IoT. Computer Networks. https://doi.org/10.1016/j.comnet.2025.110215

Liu, T., et al. (2025). Privacy-preserving federated intrusion detection in IoT. Computers & Security. https://doi.org/10.1016/j.cose.2025.103412

Du, Q., et al. (2025). Distributed federated IDS for edge IoT networks. Future Generation Computer Systems. https://doi.org/10.1016/j.future.2025.03.021

Peng, H., et al. (2025). Knowledge distillation in federated intrusion detection. Sensors, 25(7), 3311. https://doi.org/10.3390/s25073311

Tfaily, F., et al. (2025). Graph-based federated intrusion detection for IoT. Scientific Reports. https://doi.org/10.1038/s41598-025-92031-2

Ioannou, I., et al. (2025). Green federated IDS for IoT networks. Computer Communications. https://doi.org/10.1016/j.comcom.2025.01.011

Hamdy, H., et al. (2025). Self-attention federated IDS for IoT. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3357812

Taheri, R., et al. (2025). Explainable AI for federated intrusion detection. Electronics. https://doi.org/10.3390/electronics14224508

Li, S., et al. (2025). SHAP-based explainable intrusion detection in IoT. IEEE IoT Journal. https://doi.org/10.1109/JIOT.2025.3361234

Salameh, H., et al. (2025). XAI-enhanced intrusion detection for IoT. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3345122

Ghenai, A., et al. (2025). Explainable GAN-based IDS for IoT. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3339876

Labraoui, N., et al. (2025). Explainable federated IDS in fog computing. Cluster Computing. https://doi.org/10.1007/s10586-025-05211-3

Stolz, A., & Zhang, Y. (2025). Explainable federated IDS for IoT. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3342211

Sahoo, J. (2026). Explainable federated IDS for edge IoT. Information Fusion. https://doi.org/10.1016/j.inffus.2026.01.003

Villegas, W. (2025). Autonomous federated defense in IoT. Frontiers in Communications. https://doi.org/10.3389/frcmn.2025.00221

Bensaid, R. (2025). Federated IDS for smart home IoT. Cluster Computing. https://doi.org/10.1007/s10586-025-05177-2

Yang, C. (2025). Lightweight federated IDS. Computer Networks. https://doi.org/10.1016/j.comnet.2025.110101

Fraihat, S. (2026). Deep federated IDS for IoMT. Journal of Ambient Intelligence. https://doi.org/10.3233/AIS-260014

Khraisat, A. (2025). Privacy-aware federated IDS. Journal of Cloud Computing. https://doi.org/10.1186/s13677-025-00489-3

Bouzeraib, W. (2025). GAN-based federated IDS. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3334567

Deshmukh, A. (2025). Federated IDS optimization. Sensors. https://doi.org/10.3390/s25041234

Chandu, G. (2025). Distributed federated IoT security. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3349987

Gui, G. (2025). AI-based intrusion detection. Elsevier. https://doi.org/10.1016/j.comnet.2025.110310

Aburasain, R. (2025). Deep IDS optimization. Elsevier. https://doi.org/10.1016/j.future.2025.04.002

Nagaradjane, P. (2025). Energy-efficient federated IDS. Computer Communications. https://doi.org/10.1016/j.comcom.2025.02.012

Angin, P. (2025). Federated IoT security. Computer Communications. https://doi.org/10.1016/j.comcom.2025.01.017

Rodriguez, R. (2025). Federated cybersecurity infrastructure. Sensors. https://doi.org/10.3390/s25051234

Vassiliou, V. (2025). Edge-cloud federated IDS. Computer Communications. https://doi.org/10.1016/j.comcom.2025.03.018

Uribe, A. (2025). IoT cyberattack detection with FL. Academic Press. https://doi.org/10.1016/B978-0-12-823123-4.00010-3

Hans, P. (2025). Blockchain federated IDS. arXiv. https://doi.org/10.48550/arXiv.2507.11234

Alazab, M. (2025). Industrial IoT federated IDS. Discover IoT. https://doi.org/10.1007/s43926-025-00155-1

Graffi, K. (2025). Performance of federated IDS. Journal of Sensor Networks. https://doi.org/10.3390/jsn14030123

Devine, M. (2026). Explainable federated AI IDS. IEEE Transactions. https://doi.org/10.1109/TII.2026.3350123

Zeghib, N. (2025). Federated anomaly detection IoT. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3347654

Maleh, Y. (2025). Explainable IDS survey. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3345512

Ennaji, E. (2025). SHAP explainable IoT IDS. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3346671

Saez-de-Camara, U. (2025). IoT datasets for explainable IDS. Academic Press. https://doi.org/10.1016/B978-0-12-823456-3.00005-2