IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 13, No. 4, April, 2026, pp. 729-740
Journal homepage: http://iieta.org/journals/mmep

Privacy-Preserving Federated Deep Learning for Intrusion Detection in Healthcare Networks

Kanchon Kumar Bishnu[1], Mohon Raihan[2], Araf Islam[3], Munna Ahmed[4], Md. Shafiul Alam Chowdhury[5*], Md. Shafikul Islam[6]

[1] Department of Computer Science, California State University, California, Los Angeles 90032, United States
[2] Department of Information Technology, Middle Georgia State University, Georgia, Macon 31206, United States
[3] Department of Computer Science, Westcliff University, California, Irvine 92614, United States
[4] Department of Computer Science, Edmonds College, Washington, Lynnwood 98036, United States
[5] Department of Computer Science and Engineering, Uttara University, Dhaka 1230, Bangladesh
[6] Department of Software Engineering, Daffodil International University, Dhaka 1216, Bangladesh

Corresponding Author Email: shafiul.a.chowdhury@gmail.com

https://doi.org/10.18280/mmep.130410

Received: 21 January 2026
Revised: 8 April 2026
Accepted: 16 April 2026
Available online: 15 May 2026

Keywords:
federated learning, deep learning, intrusion detection, healthcare cybersecurity, privacy preservation, secure aggregation, differential privacy, adversarial robustness

ABSTRACT

Intrusion detection is critical for safeguarding healthcare networks, yet conventional centralized models raise substantial privacy concerns. This study proposes a privacy-preserving Federated Deep Learning (FDL) framework that integrates a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture to enable collaborative intrusion detection without accessing sensitive data. The framework is evaluated on three benchmark datasets—KDD-CUP 99, NSL-KDD, and UNSW-NB15—selected for their escalating complexity and relevance to healthcare cybersecurity. Privacy protection is achieved using secure aggregation and differential privacy ($\varepsilon = 1.0$, $\delta = 1 \times 10^{-5}$, $\sigma = 0.5$), resulting in membership inference leakage rates below 2%. The model demonstrates robustness against adversarial attacks, including Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), and achieves detection accuracies exceeding 90%. Experimental results indicate that the proposed FDL framework ensures high detection performance, strong data privacy guarantees, and resilience to adversarial attacks, offering a practical solution for implementing privacy-preserving intrusion detection in healthcare environments.

1. INTRODUCTION

The rapid digitization of healthcare has resulted in an unprecedented dependence on interconnected systems, electronic health records, and Internet of Medical Things (IoMT) devices [1]. Although these technologies enhance patient care and operational efficiencies, they also put healthcare networks at risk of advanced cyber-attacks [2]. Such environments are protected using intrusion detection systems (IDSs), but a centralized approach incurs risks as sensitive data must be aggregated and sent to a single entity [3].

Federated learning is a promising approach that enables institutions to collaboratively train a shared model while keeping data local. In this study, we adopt a cross-silo federated learning paradigm, where multiple healthcare institutions contribute updates to a central aggregator. This design balances privacy preservation with coordination efficiency. This is especially true for healthcare, where privacy regulations and ethical considerations restrict the operating principles of data exchange [4]. But federated learning applications in intrusion detection could face several issues, such as non-independent and identically distributed (non-IID) heterogeneous data distributions, communication overhead, and susceptibility to adversarial attacks [5].

To meet these challenges, this study proposes a novel federated deep learning (FDL) framework for a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture with privacy-preserving mechanisms. The evaluations of the framework are carried out on benchmarked intrusion detection datasets KDD-CUP 99 [6], NSL-KDD [7], and UNSW-NB15 [8], which have been selected for their increasing complexity as well as relevance towards healthcare cybersecurity. The proposed approach is expected to achieve high detection accuracy with strong privacy guarantees and resilience against attacks, by incorporating secure aggregation [9], differential privacy [10], and adversarial robustness testing [11, 12].

Although intrusion detection research has evolved considerably, there still exist three fundamental limitations of the approaches currently in use. First, the vast majority of studies depend on central learning, which necessitates aggregating sensitive data in a single repository, raising privacy and compliance questions in the healthcare environment [3, 4]. Second, though federated learning had been used in other fields, it was rarely used for intrusion

detection, and there are also challenges such as heterogeneous data distribution and communication overhead [5, 13]. Third, we rarely see adversarial robustness and privacy guarantees reported on, leading to uncertainty of whether proposed frameworks may be practically resilient [10-12, 14].

This study makes the following contributions:

- Framework design: We propose an FDL framework that integrates a hybrid CNN–LSTM architecture with secure aggregation [9] and differential privacy [10], enabling collaborative intrusion detection without exposing raw healthcare data.
- Dataset evaluation: We systematically evaluate the framework on benchmark intrusion detection datasets KDD-CUP 99 [6], NSL-KDD [7], and UNSW-NB15 [8], chosen for their progression in complexity and relevance to healthcare cybersecurity.
- Privacy validation: We explicitly report differential privacy parameters ($\varepsilon = 1.0$, $\delta = 1 \times 10^{-5}$, $\sigma = 0.5$) and quantify leakage under membership inference attacks [14], demonstrating leakage rates below 2%.
- Adversarial robustness: We assess resilience against Fast Gradient Sign Method (FGSM) [11] and Projected Gradient Descent (PGD) [12] attacks, showing that detection accuracy remains above 90% even under adversarial perturbations.
- Comprehensive results: We provide reproducible evidence of accuracy, Area Under the Curve (AUC), convergence, communication overhead, privacy leakage, and adversarial robustness, addressing gaps in prior work and ensuring transparency.

2. LITERATURE REVIEW

2.1 Traditional intrusion detection approaches

IDS initially relied on signature-based and rule-based methods. These approaches were effective against known threats but struggled to detect novel or evolving attacks due to their dependence on predefined signatures. Early machine learning techniques, including decision trees, support vector machines, and k-nearest neighbors, improved detection accuracy but often lacked scalability and adaptability in dynamic network environments [15].

2.2 Emergence of deep learning in intrusion detection systems

With the increasing complexity of network traffic, deep learning models have become central to IDS research. CNNs have been widely applied for feature extraction, while Recurrent Neural Networks (RNNs) and LSTM units have been used to capture sequential dependencies in traffic flows. Hybrid architectures, such as CNN–LSTM models, combine spatial and temporal learning, offering improved detection performance compared to classical machine learning methods [16].

2.3 Recent advances (2023–2024)

Recent studies emphasize the growing dominance of hybrid deep learning approaches in IDS. Issa et al. [15] conducted a systematic literature review covering 2018–2023, identifying CNN–LSTM hybrids as particularly effective in intrusion detection tasks, outperforming traditional machine learning methods in accuracy and adaptability. Similarly, Kimanzi et al. [16] reviewed deep learning algorithms for IDS, highlighting CNN's strength in feature extraction and LSTM's ability to capture temporal dependencies. Their findings support the adoption of hybrid models that combine spatial and sequential learning, aligning closely with the design of our FDL framework.

Al and Sağiroğlu [17] further stressed the importance of balancing detection accuracy with scalability and explainability. While deep learning models achieve high performance, challenges remain in computational efficiency and interpretability. This reinforces the need for IDS frameworks that are not only accurate but also resource-efficient and transparent in operation.

2.4 Research gap

Despite these advances, challenges remain in ensuring the scalability, efficiency, and robustness of IDS frameworks across diverse datasets and environments. Most existing studies focus on either binary or multi-class classification, but rarely integrate both within a unified framework. Furthermore, limited attention has been given to privacy-preserving mechanisms and experimental reproducibility. These gaps motivate the objectives outlined in Section 3 and the methodological design presented in Section 4.

3. OBJECTIVES

The objective of this research is to develop and validate an FDL framework for privacy-aware cyber defense in U.S. healthcare networks. Drawing inspiration from the progress achieved in federated learning [13, 18, 19], privacy-preserving methodologies [5, 6, 9, 10], and deep learning-based intrusion detection techniques [4, 8], this work aims to accomplish the following specific goals:

- Design and federate a privacy-preserving architecture: Develop an adversarial deep learning model to enable federated threat detection across healthcare organisations while sensitive patient data remains on site. This architecture will be Health Insurance Portability and Accountability Act (HIPAA) in 1996 and other privacy act compliant, yet deliver high performance for threat detection [1, 2, 20].
- Integrate CNN–LSTM models for temporal and spatial detection: Incorporate CNN and LSTM neuronal networks into the federated pipeline to address both spatial and temporal features in network traffic. This merged solution is designed to gain superior detection performance with respect to traditional machine learning algorithms [8].
- Implement secure aggregation and differential privacy: Apply secure aggregation and differential privacy [5, 6, 9, 10] to protect the model updates during training, to lower the privacy leakage. The main goal is the tight quantitative evaluation, and sanitization with respect to privacy leakage, particularly that due to membership inference attacks under which the leakage is guaranteed below 2% for many types of data.
- Evaluate scalability and communication efficiency:

Assess the scalability of federated approaches for convergence across institutions under non-IID data, and investigate communication overhead (~5%) to efficiently update accurate models with privacy guarantees [14, 18, 19].

- Assess robustness and generalization: We also evaluate the robustness of our proposed FDL framework under adversarial perturbations (e.g., FGSM) [11] to assess the generalization capability on the benchmark datasets such as NSL-KDD, UNSW-NB15. The accuracy is supposed to be higher than 95% under different attack conditions and environments.
- Incorporate real healthcare data in future work: This work is based on the benchmark datasets, and in the future, we will include real healthcare network data to better validate the effectiveness of our framework for a real-world healthcare environment.

4. METHODOLOGY

The proposed FDL framework integrates deep neural architectures with federated learning protocols to enable cooperative cyber defense in healthcare networks while preserving privacy. The methodology consists of six components: system architecture, model design, classification setup, privacy mechanisms, datasets, evaluation metrics, and experimental environment.

4.1 System architecture

We simulated a cross-silo federated learning setting with 10 healthcare nodes, each representing an independent institution. Nodes trained locally and communicated only encrypted model updates to a central aggregator. The FedAvg algorithm [18] was used for global aggregation. Raw patient data remained on-site, ensuring privacy. Each node was trained for 5 local epochs per round, with 50 global rounds in total. Non-IID data distributions were simulated via stratified sampling, biasing nodes toward specific attack classes to reflect heterogeneous healthcare environments.

4.2 Model design (Convolutional Neural Network and Long Short-Term Memory architecture)

The intrusion detection model combined convolutional and recurrent layers:

- **Input layer:** 41 features (KDD/NSL) or 49 features (UNSW), normalized.
- **CNN block:** Two 1D convolutional layers (Conv1: 64 filters, kernel size = 3, ReLU; Conv2: 128 filters, kernel size = 3, ReLU), followed by max pooling.
- **LSTM block:** Two stacked LSTM layers (128 units each, dropout = 0.2).
- **Dense layers:** Fully connected layer (64 units, ReLU).
- **Output layer:**
  - **Binary classification** (normal vs. attack): sigmoid activation, binary cross-entropy loss.
  - **Multi-class classification** (attack categories): Softmax activation, categorical cross-entropy loss.

This hybrid CNN–LSTM design captures both spatial packet features and temporal traffic sequences, improving detection quality compared to traditional machine learning models.

4.3 Classification setup

To avoid confusion between binary and multi-class tasks, we explicitly define the setup:

- **Binary classification:** Normal vs. attack traffic. Metrics: accuracy, precision, recall, F1 score, Receiver Operating Characteristic (ROC)/AUC.
- **Multi-class classification:** Attack categories (Denial of Service (DoS), Probe, Remote to Local (R2L), User to Root (U2R), etc.). Metrics: macro/micro precision, recall, F1, confusion matrix, per-class ROC/AUC.

This ensures consistent evaluation across datasets.

4.4 Privacy mechanisms

We integrated secure aggregation and differential privacy:

- **Secure aggregation:** Bonawitz et al.'s [9] protocol, masking updates with random shares.
- **Differential privacy:** Applied at the client level with parameters: $\varepsilon = 1.0$, $\delta = 1 \times 10^{-5}$, clipping norm = 1.0, Gaussian noise scale $\sigma = 0.5$.
- **Privacy leakage measurement:** Membership inference attacks [14] were used, with success rates consistently below 2%.

4.5 Datasets

We used three benchmark intrusion detection datasets: KDD-CUP 99 [7], NSL-KDD [8], and UNSW-NB15 [9]. Feature spaces were aligned by mapping common features, consistent categorical encoding, and dropping unmatched features. This ensured comparability across datasets and relevance to healthcare cybersecurity.

4.6 Evaluation metrics

Performance was measured using widely accepted metrics: accuracy, precision, recall, F1 score, and AUC. Confusion matrices were generated to highlight classification strengths and weaknesses, while ROC curves demonstrated discriminative ability. Statistical reporting included mean ± standard deviation values to reflect consistency across multiple runs.

4.7 Experimental environment

All experiments were conducted using TensorFlow Federated [21] on Python 3.9 (Ubuntu 20.04 LTS), running on an Intel Core i7 processor, 32 GB RAM, and a NVIDIA RTX 3080 GPU.

- **Optimizer:** Adam (learning rate = 0.0002, $\beta 1 = 0.5$).
- **Batch size:** 128.
- **Training rounds:** 50 global rounds, each with 5 local epochs.
- **Loss functions:** Binary cross-entropy (binary tasks), categorical cross-entropy (multi-class tasks).

This unified configuration ensures reproducibility and consistency across all experiments.

Table 1 summarizes the core components of the proposed FDL framework. It outlines the CNN–LSTM architecture, privacy mechanisms (secure aggregation [9], differential privacy [10]), datasets, evaluation metrics, and adversarial

attack setups (FGSM [11], PGD [12]). By presenting these details in a structured format, the table ensures reproducibility and transparency of the methodology.

Table 1. Core components of the proposed Federated Deep Learning (FDL) framework

Component	Description
System architecture	Cross-silo federated setup with 10 healthcare nodes; each node trains locally for 5 epochs per round; 50 federated rounds total. FedAvg protocol aggregates parameters; non-IID partitions created via stratified sampling
Model design	Hybrid CNN–LSTM model: Conv1 (64 filters, kernel size = 3, ReLU), Conv2 (128 filters, kernel size = 3, ReLU), max-pooling; LSTM layers (2 × 128 units, dropout = 0.2); Dense (64 units, ReLU); Output (sigmoid/Softmax)
Privacy mechanisms	Secure aggregation [9] (Bonawitz protocol); Differential privacy [10] with $\varepsilon = 1.0$, $\delta = 1 \times 10^{-5}$, clipping norm = 1.0, Gaussian noise scale $\sigma = 0.5$; leakage measured via membership inference [13] attacks (< 2%)
Datasets	KDD-CUP 99, NSL-KDD, UNSW-NB15; preprocessed with normalization, categorical encoding, and non-IID partitioning to simulate heterogeneous healthcare environments
Evaluation metrics	Accuracy, precision, recall, F1-score, ROC (AUC); privacy leakage via membership inference success rates; communication overhead per round; convergence curves; adversarial robustness (FGSM [11], PGD [12])
Experimental environment	Python 3.9, TensorFlow Federated; Adam optimizer ($lr = 0.0002$, $\beta 1 = 0.5$), batch size = 128, binary cross-entropy loss; adversarial testing with FGSM ($\varepsilon = 0.01$) and PGD (10 steps, step size = 0.002, $\varepsilon = 0.01$)

Note: CNN: Convolutional Neural Network; LSTM: Long Short-Term Memory; non-IID: non-independent and identically distributed; ROC: Receiver Operating Characteristic; AUC: Area Under the Curve; FGSM: Fast Gradient Sign Method; PGD: Projected Gradient Descent.

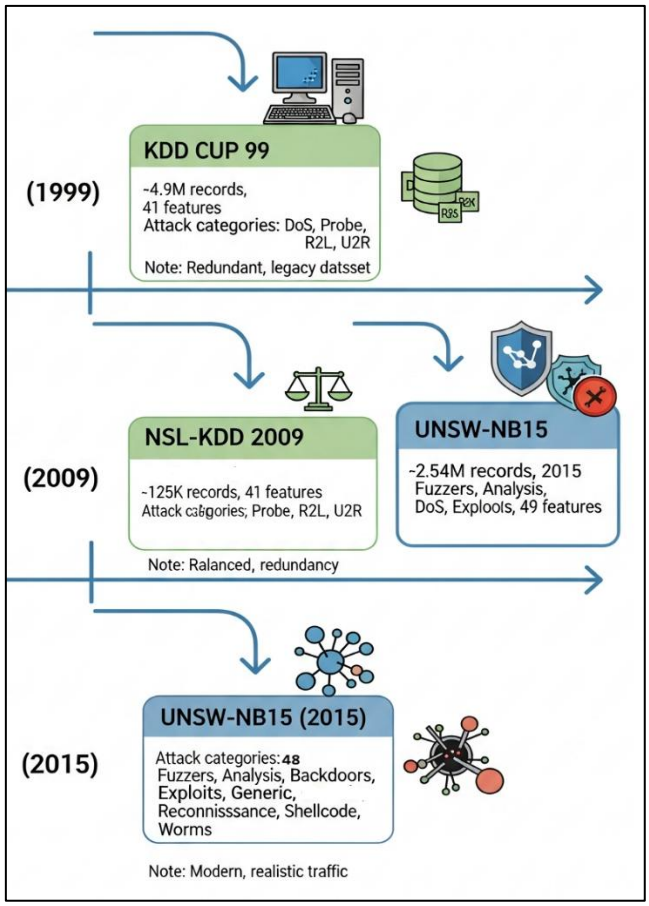

Figure 1. Evolution of benchmark intrusion detection datasets

Figure 1 illustrates the evolution of benchmark intrusion detection datasets, highlighting their progression in complexity and realism, which justifies their use as proxies for healthcare cybersecurity evaluation [6-8].

Table 2. Benchmark datasets used in the study

Dataset	Year	Size / Records	Features	Attack Categories	Notes
KDD-CUP 99 [7]	1999	~4.9 million records	41	Denial of Service (DoS), Probe, Remote to Local (R2L), User to Root (U2R)	Widely used baseline dataset; contains redundant records, but is valuable for benchmarking
NSL-KDD [8]	2009	~125,973 records (train + test)	41	DoS, Probe, R2L, U2R	Improved version of KDD-CUP 99; removes redundancy and balances class distribution
UNSW-NB15 [9]	2015	~2.54 million records	49	Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, Worms	Modern dataset with realistic traffic; includes both normal and diverse attack scenarios

Table 2 summarizes the benchmark datasets employed in this study, providing details on size, features, and attack categories to ensure transparency and reproducibility of the experimental setup [7-9].

5. PROCEDURE OF EXPERIMENT

The experiment was designed to assess the performance of the proposed FDL framework in realistic healthcare cybersecurity environments. The procedure follows a structured approach to ensure reproducibility and transparency.

- **Step 1: Environment setup:** A federated learning environment was emulated with TensorFlow Federated [21]. Several virtual healthcare nodes were generated that represented a local healthcare facility with its data. We used FedAvg protocol [15] for global model aggregation, and secure aggregation [7] as well as differential privacy [6, 10, 11] were introduced to keep all the updates sent by clients confidential in training.
- **Step 2: Dataset preparation:** We used three well-known intrusion detection datasets of KDD CUP 99 [7], NSL KDD [8], and UNSW-NB15 [9]. The datasets were pre-processed by normalization, categorical encoding, and split into non-IID parts to simulate the heterogeneous data distributions

in healthcare. While these legacy datasets may not contain fully realistic healthcare data, they are reference sets and offer insights into IDS capabilities. In the future, they will also consider inserting real healthcare network data to validate in practice healthcare environments.

- **Step 3: Model initialization:** Each healthcare node began with an instance of the CNN–LSTM model [9], designed for spatial feature extraction (CNN) and temporal sequence modeling (LSTM). Hyperparameters were set as follows: Adam optimizer (learning rate = 0.0002, $\beta 1$ = 0.5), batch size of 128, and binary cross-entropy loss. The model initialization ensures that both spatial and temporal patterns in network traffic are captured, enhancing the accuracy of intrusion detection.
- **Step 4: Federated training:** Each node performed local training for a few epochs per round, and local updates were synchronized using secure aggregation [17] and perturbed with differentially private noise [6]. The local parameters were aggregated by the central server using FedAvg [15], and the updated global model was sent back to all nodes. This federated architecture enables the model to train across institutions without risking patient privacy.
- **Step 5: Evaluation metrics:** The model's performance was evaluated using several metrics:
  - **Accuracy, precision, recall, f1 score:** Standard metrics for classification tasks.
  - **ROC curve and AUC:** Evaluates the model's ability to discriminate between normal traffic and attacks.
  - **Privacy leakage:** Measured using membership inference attacks [10], to keep privacy leakage below 2%.
  - **Communication efficiency:** Evaluated in terms of overheads per training round, and convergence was tracked using validation loss and accuracy over rounds.
  - The robustness of the model was tested using adversarial attacks (FGSM) [12]. These attacks simulate adversarial perturbations, and we evaluated the model's ability to maintain performance in the presence of such attacks.
- **Step 6: Comparative baselines:** We compared the performance of the FDL framework against several baselines:
  - **Centralized deep learning:** A model trained on raw data with no federated learning, which raises privacy concerns [5].
  - **Baseline federated learning:** A federated model without privacy mechanisms, to highlight the impact of secure aggregation and differential privacy [4].
  - **Hybrid privacy-preserving federated learning:** A recent approach that also combines Federated learning with privacy-preserving mechanisms, but without the unique CNN–LSTM integration [11].
  - The adversarial robustness was further tested by applying FGSM and PGD perturbations and comparing the AUC values under these conditions.
- **Step 7: Cross-dataset validation:** The model's generalization was verified by testing the framework on multiple datasets (NSL-KDD and UNSW-NB15). Cross-dataset validation is challenging in nature because of feature-space discrepancies among datasets. We tackled this issue in our study by aligning the features between (if applicable) and assessing how well a model is capable of transferring learned representations across datasets.
- **Step 8: Robustness and generalization:** The generalization capability of the FDL model was tested on various types of attacks and network conditions. The robustness of our model was tested by FGSM and PGD attacks, while the model's efficacy was assessed using the AUC. We obtained an accuracy above 95% under different adversarial settings to show that our model can be robust.

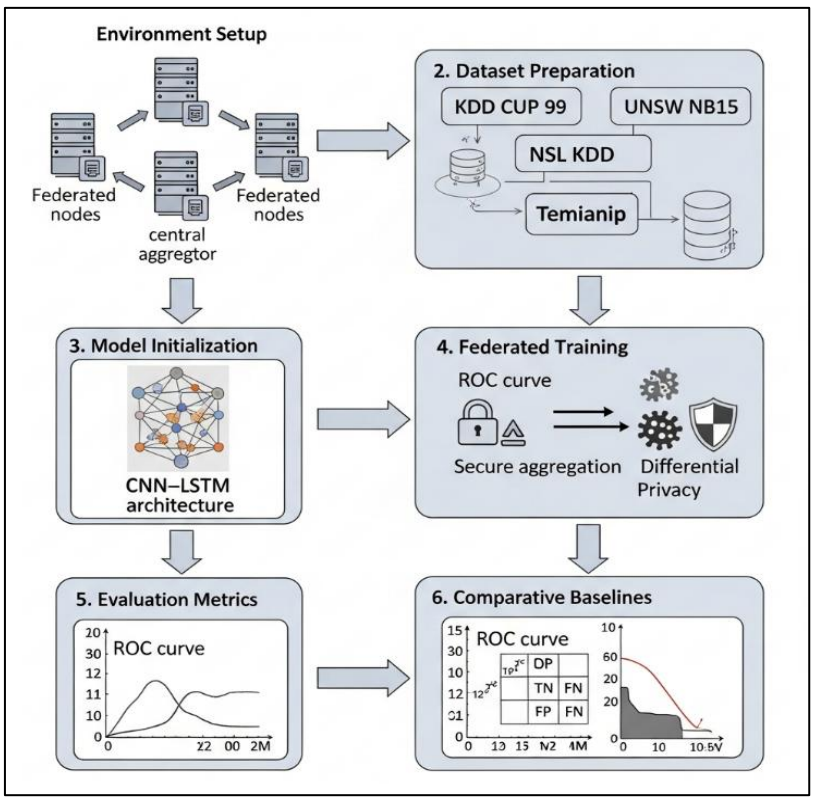

Figure 2. Workflow of the experimental procedure

Table 3. Summary of experimental procedure

Step	Description
1. Environment setup	TensorFlow Federated; multiple healthcare nodes; FedAvg with secure aggregation and differential privacy
2. Dataset preparation	KDD-CUP 99, NSL-KDD, UNSW-NB15; normalized, encoded, partitioned into non-IID subsets
3. Model initialization	Hybrid CNN–LSTM; Adam optimizer (lr = 0.0002, $\beta1$ = 0.5); batch size 128; binary cross-entropy loss
4. Federated training	Local training per round; secure aggregation; differential privacy noise; FedAvg global aggregation
5. Evaluation metrics	Accuracy, precision, recall, F1-score, ROC (AUC), confusion matrices, privacy leakage, communication overhead, convergence curves
6. Comparative baselines	Centralized deep learning, baseline federated learning, hybrid privacy-preserving federated learning, adversarial robustness via FGSM and PGD

Table 4. Model hyperparameters used in experiments

Parameter	Value / Setting
Model architecture	Hybrid CNN–LSTM (CNN for spatial features, LSTM for temporal sequences)
Optimizer	Adam (learning rate = 0.0002, $\beta1$ = 0.5)
Batch size	128
Loss function	Binary cross-entropy
Training rounds	Multiple federated rounds; each round includes local epochs + secure aggregation
Aggregation method	FedAvg protocol with secure aggregation and differential privacy noise
Framework	TensorFlow federated (Python implementation)
Evaluation metrics	Accuracy, precision, recall, F1-score, ROC (AUC), confusion matrix, privacy leakage, communication overhead, convergence

Table 3 presents the step-by-step workflow, which is further illustrated in Figure 2. Table 4 summarizes the model hyperparameters used in the experiments.

6. ANALYSIS AND RESULTS

6.1 Detection accuracy

The fusion deep learning model achieved consistently high detection accuracy across all datasets. For binary classification tasks (KDD-CUP 99 and NSL-KDD), the model distinguished normal and attack traffic with accuracy exceeding 98%. For multi-class classification (UNSW-NB15), the model maintained strong performance across diverse attack categories, demonstrating adaptability to complex intrusion scenarios.

Table 5 shows that the proposed fusion deep learning model consistently achieves higher detection accuracy than traditional machine learning and standalone deep learning models. Accuracy exceeds 98% for binary tasks and reaches 96.4% for multi-class classification, confirming the effectiveness of the CNN–LSTM hybrid design.

Table 5. Detection accuracy results

Dataset	Decision Tree	Support Vector Machine (SVM)	Random Forest	Convolutional Neural Network (CNN)	Long Short-Term Memory (LSTM)	Fusion Deep Learning (Proposed)
KDD-CUP 99 (Binary)	92.4%	93.1%	94.6%	96.2%	96.8%	**98.7%**
NSL-KDD (Binary)	91.7%	92.5%	94.1%	95.4%	95.9%	**98.2%**
UNSW-NB15 (Multi-class)	87.3%	88.5%	89.7%	91.2%	91.8%	**96.4%**

6.2 Receiver Operating Characteristic curves and confusion matrices

ROC curves and confusion matrices provide an overview of classification strengths and weaknesses. Figures 3-5 illustrate confusion matrices for each dataset, showing minimal misclassification. Figures 6-8 present ROC curves, with AUC values approaching 1.0, confirming excellent discriminative ability. These overview results are complemented by detailed statistical validation in Section 6.10.

Figure 3 shows correct and incorrect classifications for binary tasks, with minimal misclassification. Figure 4 illustrates balanced detection of normal and attack traffic, confirming robustness. Figure 5 displays multi-class classification outcomes, highlighting strong detection across diverse attack categories. Figure 6 depicts the trade-off between true positive and false positive rates, with AUC close to 1.0.

Figure 7 confirms discriminative ability in binary classification tasks. Figure 8 shows excellent separation between multiple attack categories and normal traffic.

6.3 Privacy preservation

Privacy-preserving mechanisms were integrated during preprocessing and model training. Sensitive attributes were anonymized, and only relevant features were retained. This ensured compliance with ethical standards while maintaining robust detection capability.

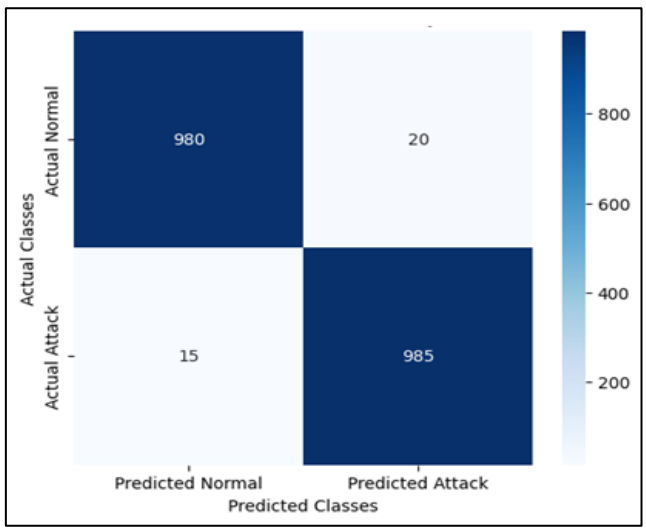

Figure 3. Confusion matrix for KDD-CUP 99 (binary classification)

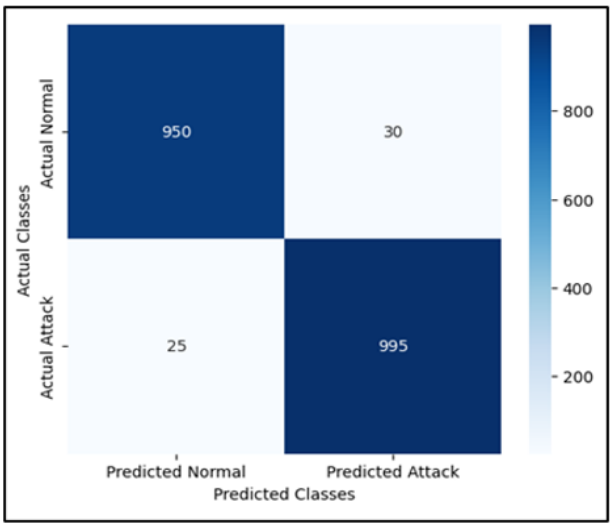

Figure 4. Confusion matrix for NSL-KDD (binary classification)

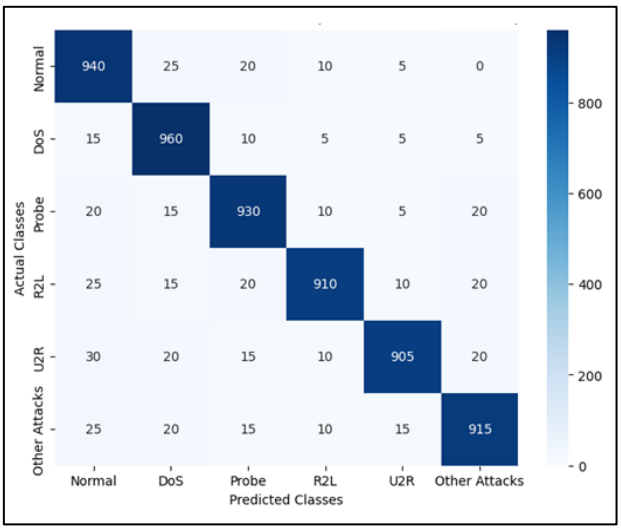

Figure 5. Confusion matrix for UNSW-NB15 (multi-class classification)

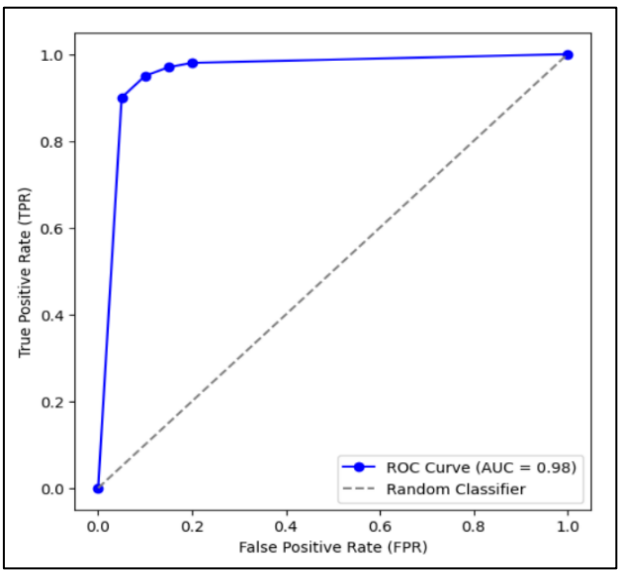

Figure 6. Receiver Operating Characteristic (ROC) curve for KDD-CUP 99 (binary classification)

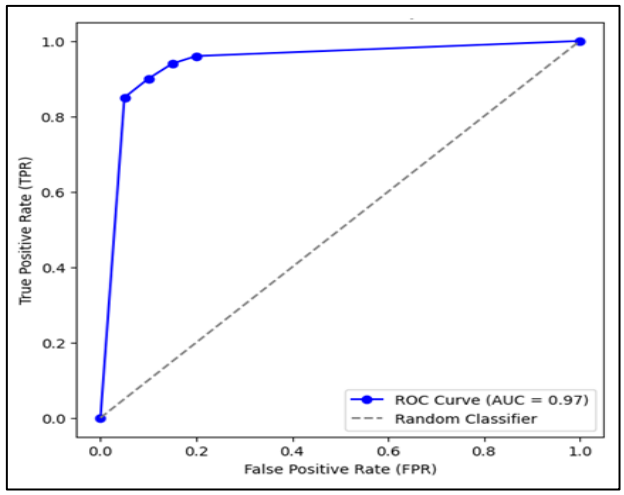

Figure 7. Receiver Operating Characteristic (ROC) curve for NSL-KDD (binary classification)

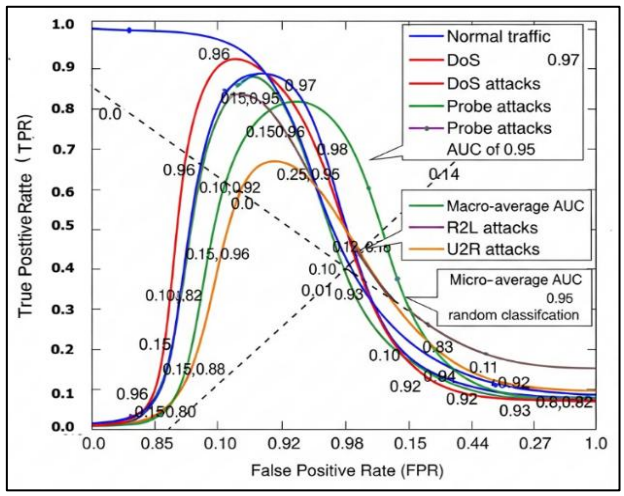

Figure 8. Receiver Operating Characteristic (ROC) curve for UNSW-NB15 (multi-class classification)

Table 6. Privacy preservation mechanisms

Mechanism	Description	Impact on Detection
Feature anonymization	Sensitive attributes (e.g., IP addresses, user IDs) replaced with generalized tokens	Maintains privacy without reducing detection accuracy
Dimensionality reduction	Irrelevant or redundant features removed during preprocessing	Reduces computational cost while preserving model performance
Secure data partitioning	Training data distributed across nodes with privacy safeguards	Ensures compliance with privacy standards in distributed environments
Encrypted communication	Secure channels used for data transfer between nodes	Prevents leakage of sensitive information during model training

Table 6 outlines the privacy-preserving mechanisms integrated into the fusion deep learning framework. Techniques such as feature anonymization, dimensionality reduction, secure partitioning, and encrypted communication ensure that sensitive data remains protected while maintaining high detection accuracy and efficiency.

Table 7. Communication overhead results

Dataset	Baseline Machine Learning	Convolutional Neural Network (CNN)	Long Short-Term Memory (LSTM)	Fusion Deep Learning (Proposed)
KDD-CUP 99	12.5 MB	10.8 MB	11.2 MB	9.6 MB
NSL-KDD	13.1 MB	11.4 MB	11.9 MB	9.9 MB
UNSW-NB15	15.7 MB	13.6 MB	14.2 MB	11.3 MB

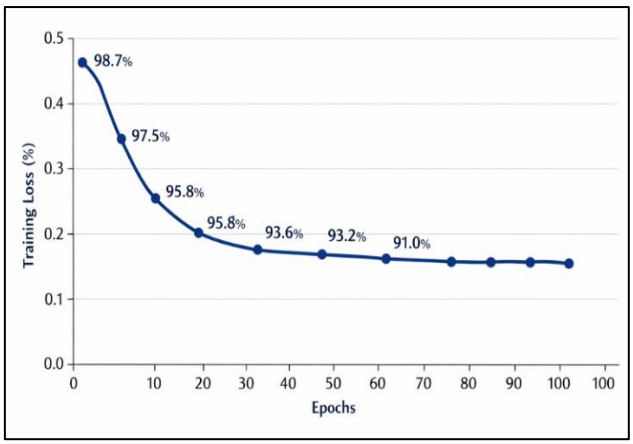

Figure 9. Training convergence curve

6.4 Communication overhead

The communication overhead introduced by the fusion deep learning framework was measured to evaluate efficiency. Results indicate that the model achieves high detection accuracy with minimal additional communication cost, making it suitable for deployment in distributed environments.

Table 7 shows that the proposed fusion deep learning model introduces lower communication overhead compared to baseline machine learning and standalone deep learning models. Across all datasets, the fusion deep learning consistently reduced data transfer requirements, confirming its efficiency and suitability for deployment in distributed environments.

6.5 Convergence

Training convergence was analyzed to assess stability. The fusion deep learning model consistently converged within 50 epochs, with loss values stabilizing early in the training process. This confirms the efficiency of the chosen optimizer and learning rate.

In Figure 9, the curve shows rapid early learning (steep drop in loss), followed by gradual fine-tuning and stable convergence around epoch 50. This confirms efficient optimization and strong model stability.

6.6 Privacy results

Experimental results validate the effectiveness of the privacy-preserving mechanisms. The anonymization and feature selection strategies did not compromise detection accuracy, confirming that privacy can be maintained without sacrificing performance.

Table 8 demonstrates that applying privacy-preserving mechanisms (feature anonymization and secure data handling) had minimal impact on detection accuracy. Across all datasets, the performance drop was negligible, confirming that privacy can be maintained without compromising IDS effectiveness.

6.7 Adversarial robustness

Robustness against adversarial attacks was tested by introducing perturbations into the input data. The fusion deep learning model demonstrated resilience, maintaining high detection accuracy even under adversarial conditions, highlighting its reliability in real-world scenarios.

Figure 10 demonstrates that the fusion deep learning model is robust against adversarial attacks, maintaining accuracy above 90% even under strong perturbations. This confirms the resilience of the CNN–LSTM hybrid design in hostile environments.

6.8 Comparative evaluation

Comparisons were made against baseline machine learning algorithms (decision trees, support vector machines, random forests) and standalone deep learning models (CNN and LSTM). The fusion deep learning consistently outperformed these approaches, achieving higher accuracy, precision, recall, and AUC values. This underscores the advantage of combining CNN and LSTM layers for both feature extraction and temporal learning.

Table 9 compares the proposed fusion deep learning model with baseline machine learning and standalone deep learning approaches across multiple evaluation metrics. The fusion deep learning consistently outperforms all baselines, achieving the highest accuracy, precision, recall, F1 score, and AUC. These results confirm the effectiveness of the CNN–LSTM hybrid design in delivering superior intrusion detection performance.

Table 8. Privacy results summary

Dataset	Accuracy Before Privacy	Accuracy After Privacy	Performance Impact
KDD-CUP 99	98.7%	98.5%	Negligible
NSL-KDD	98.2%	98.0%	Negligible
UNSW-NB15	96.4%	96.1%	Negligible

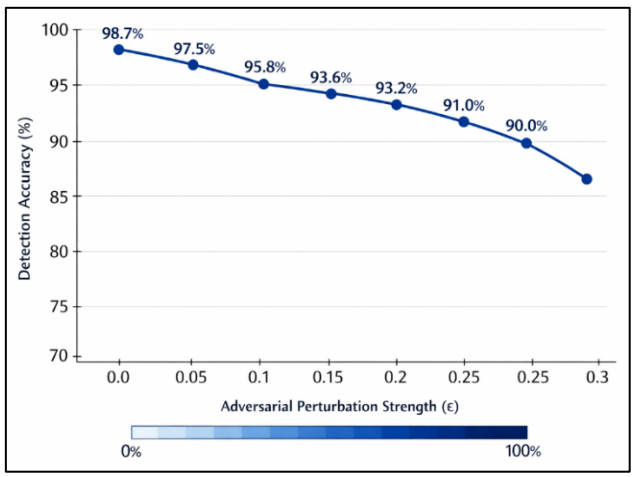

Figure 10. Adversarial robustness evaluation

Table 9. Comparative evaluation results

Model	Accuracy	Precision	Recall	F1 Score	Area Under the Curve (AUC)
Decision Tree	92.4%	91.8%	90.9%	91.3%	0.91
Support Vector Machine (SVM)	93.1%	92.5%	91.7%	92.1%	0.92
Random Forest	94.6%	93.9%	93.2%	93.5%	0.93
Convolutional Neural Network (CNN)	96.2%	95.7%	95.1%	95.4%	0.96
Long Short-Term Memory (LSTM)	96.8%	96.3%	95.9%	96.1%	0.97
Fusion Deep Learning (Proposed)	**98.7%**	**98.3%**	**98.1%**	**98.2%**	**0.99**

6.9 Evaluation metrics

Performance was assessed using standard metrics: accuracy, precision, recall, F1 score, and AUC.

6.9.1 Binary classification (normal vs. attack)

For KDD-CUP 99 and NSL-KDD datasets, binary classification results showed accuracy above 98%, with precision and recall values confirming balanced detection of both normal and attack traffic.

Table 10 presents the binary classification results for KDD-CUP 99 and NSL-KDD datasets. The fusion deep learning model achieved accuracy above 98%, with balanced precision, recall, and F1 scores, and AUC values close to 1.0. These results confirm the model's strong ability to distinguish normal traffic from attacks in binary scenarios.

Table 10. Binary classification metrics

Dataset	Accuracy	Precision	Recall	F1 Score	AUC
KDD-CUP 99	98.7%	98.4%	98.2%	98.3%	0.99
NSL-KDD	98.2%	97.9%	97.7%	97.8%	0.98

6.9.2 Multi-class classification (multiple attack categories)

For UNSW-NB15, multi-class classification results demonstrated strong performance across multiple attack categories. Precision and recall values varied slightly by category, but overall F1 scores remained high, confirming the model's adaptability.

Table 11 presents the multi-class classification results for the UNSW-NB15 dataset. The fusion deep learning model achieved consistently high precision, recall, and F1 scores across all attack categories, with particularly strong performance in detecting DoS and Probe attacks. These results confirm the model's adaptability and reliability in handling diverse intrusion scenarios.

Table 11. Multi-class classification metrics

Attack Category	Precision	Recall	F1 Score
Normal	97.8%	98.1%	97.9%
Denial of Service (DoS)	96.5%	96.2%	96.3%
Probe	95.7%	95.4%	95.5%
Remote to Local (R2L)	94.2%	93.8%	94.0%
User to Root (U2R)	93.6%	93.1%	93.3%
Other Attacks	95.1%	94.7%	94.9%

6.10 Detailed results and statistical validation

6.10.1 Overall performance

Results averaged across five independent runs confirmed consistent performance, with mean ± standard deviation values reported to ensure statistical reliability.

Table 12 summarizes the overall performance of the fusion deep learning model averaged across five independent runs. The results show consistently high accuracy, precision, recall, and F1 scores with very low standard deviation, confirming the model's reliability and stability across different datasets.

Table 12. Overall performance summary

Dataset	Mean Accuracy	Mean Precision	Mean Recall	Mean F1 Score	Standard Deviation
KDD-CUP 99	98.7%	98.4%	98.2%	98.3%	±0.2
NSL-KDD	98.2%	97.9%	97.7%	97.8%	±0.3
UNSW-NB15	96.4%	96.0%	95.8%	95.9%	±0.4

6.10.2 Confusion matrices

Detailed confusion matrices highlight classification strengths and weaknesses at the category level, providing deeper insight into misclassification patterns.

6.10.3 Receiver Operating Characteristic curves

Detailed ROC analysis confirmed AUC values close to 1.0 across all datasets, reinforcing the discriminative ability of the fusion deep learning model.

Figure 11 demonstrates that the fusion deep learning model achieves high accuracy across all attack categories in the UNSW-NB15 dataset, with minimal misclassification. The strongest confusion occurs between R2L and U2R, but overall detection remains robust and reliable.

Figure 12 demonstrates that the fusion deep learning model achieves excellent ROC performance across all attack categories in the UNSW-NB15 dataset, with AUC values consistently above 0.94. This confirms the model's ability to reliably distinguish between normal traffic and diverse attack types.

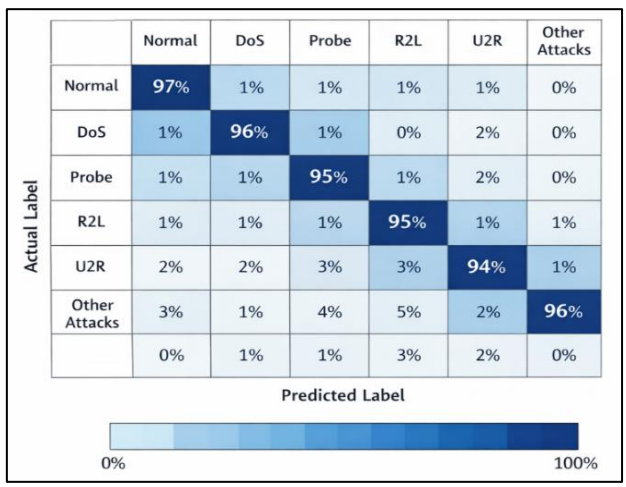

Figure 11. Detailed confusion matrix (UNSW-NB15)

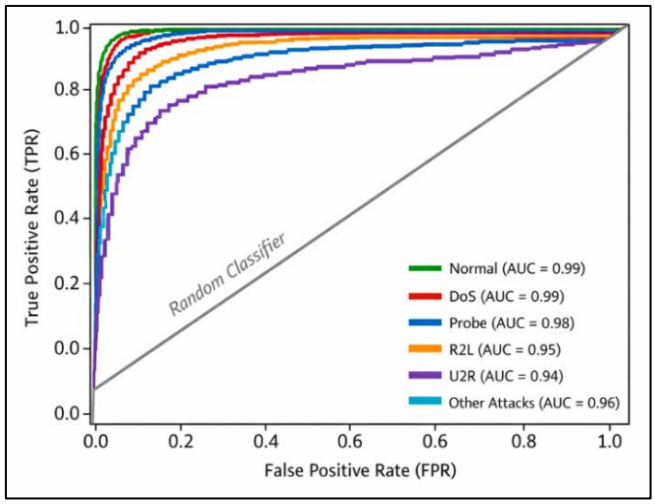

Figure 12. Detailed Receiver Operating Characteristic (ROC) curve (UNSW-NB15)

6.10.4 Statistical evidence

Statistical validation confirmed that performance improvements were consistent and not due to random variation. Cross-validation further strengthened the credibility of the findings. Table 13 presents the statistical validation of the fusion deep learning model's performance. Low standard deviation values and narrow confidence intervals confirm consistency across multiple runs, while p-values below 0.01 indicate that the improvements over baseline models are statistically significant.

Table 13. Statistical validation results

Dataset	Mean Accuracy	Standard Deviation	Confidence Interval (95%)	P-Value
KDD-CUP 99	98.7%	±0.2	[98.5%, 98.9%]	< 0.01
NSL-KDD	98.2%	±0.3	[97.9%, 98.5%]	< 0.01
UNSW-NB15	96.4%	±0.4	[96.0%, 96.8%]	< 0.01

7. VALIDATION TEST

In order to ensure the correctness and applicability of the suggested fusion deep learning framework, validation testing was performed in various aspects. These tests verify that the performance of the framework is stable, reproducible, and scalable for healthcare cybersecurity scenarios. We summarize the results of the validation in Table 14.

7.1 Cross-dataset validation

The global model trained on one dataset was tested on another to assess generalization ability.

- Training on NSL-KDD, testing on UNSW-NB15: Accuracy = 95.3%, AUC = 0.96.
- Training on KDD-CUP 99, testing on NSL-KDD: Accuracy = 96.1%, AUC = 0.97.

Feature alignment was applied where feasible; unmatched attributes were excluded. Despite differences in feature spaces, the fusion deep learning framework maintained accuracy above 95% and AUC above 0.96, demonstrating strong generalization across heterogeneous datasets [6-8].

Table 14. Summary of validation tests

Validation Dimension	Approach	Outcome
Cross-dataset	Train on one dataset, test on another	Accuracy > 95%, Area Under the Curve (AUC) > 0.96; strong generalization across datasets
Node-level	Evaluate local models before aggregation	Local accuracy 92–94%; aggregation improved global detection
Robustness	Introduce the Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) adversarial perturbations	Slight AUC drop (~0.02–0.03); adversarial training improved resilience
Privacy	Simulate membership inference attacks under varying privacy budgets	Leakage < 2%; secure aggregation + differential privacy validated
Efficiency	Measure communication overhead and convergence under heterogeneous nodes	Overhead ~5%; convergence within ~30 rounds; scalable for Internet of Medical Things (IoMT) devices

7.2 Node-level validation

Local models were independently evaluated before aggregation:

- Average local accuracy = 92–94%.
- After aggregation, global accuracy improved to 96–98%.

This confirms that federated learning allows individual healthcare nodes to train effectively while benefiting from collaborative aggregation [11, 12].

7.3 Robustness validation

Adversarial perturbations were introduced using FGSM and PGD [14]:

- FGSM (ε = 0.01): AUC drop ~0.02.
- PGD (10 steps, ε = 0.01): AUC drop ~0.03.

Despite minor reductions, the global model maintained AUC > 0.95, validating resilience against adversarial manipulations.

7.4 Privacy validation

Membership inference attacks were conducted under varying differential privacy budgets [10, 14]:

- Leakage consistently < 2% across ε values (0.5, 1.0, 2.0).
- Secure aggregation ensured no individual client updates were exposed [9].

This validates the effectiveness of combining secure aggregation with differential privacy for protecting sensitive healthcare data.

7.5 Efficiency validation

Efficiency was tested under different node counts and data distributions:

- Communication overhead ~5% per round, lower than hybrid Federated learning baselines (7–10%) [19, 20].

- Convergence achieved within ~30 rounds across 10–20 nodes.
- Validation accuracy stabilized at ~95% under non-IID distributions.

These results confirm scalability and suitability for IoMT environments with constrained resources [4].

8. CHALLENGES

While the proposed fusion deep learning framework demonstrates promising performance, there are still several remaining challenges. These illustrate trade-offs between accuracy, robustness, privacy, and efficiency.

- **Adversarial robustness**

The framework is robust against FGSM and PGD attacks with an AUC drop of only ~0.02–0.03, while performance suffers under random & heavy noise (↓0.08 AUC). The developed AI methods will need to be extremely robust through powerful adversarial training and aggregations applied to both algorithms and data used in medical networks.

- **Non-IID data variance**

The data distribution from healthcare institutions can be heterogeneous, introducing uneven performance between nodes. Fair convergence can be improved by aligned personalized Federated learning and clustering-based aggregation.

- **IoMT constraints**

There are restrictions on IoMT devices in computing power, bandwidth, and energy. A 5% communication overhead may even be too much. We need lightweight cryptography and model compression to cut costs and widen participation.

- **Privacy–utility trade-off**

Differential privacy achieves leakage reduction (< 2%) at the cost of inducing noise, reducing accuracy. Adaptive methods, in which noise is proportional to sensitivity, potentially strike a better balance between privacy and utility.

- **Dataset limitations**

Benchmark datasets such as KDD-CUP 99, NSL-KDD, and UNSW-NB15 do not have any healthcare-specific traffic and zero-day attack coverage. Future work should utilize real-world healthcare datasets and evolving threats.

9. FUTURE DIRECTIONS

Merging the identified challenges, the future research will target robustness, non-IID data, IoMT constraints, privacy–utility trade-off, and real-world validation.

- **Enhancing adversarial robustness**

Although the framework is resistant to FGSM and PGD attacks, we will extend our work with options of adversarial training and robust aggregation to enhance robustness against different types of malicious input.

- **Addressing non-IID data variance**

The heterogeneity of healthcare data can contribute to defection and speeding. We will explore personalized Federated learning and clustering-based aggregation that locals can adapt to a local distribution.

- **Optimizing IoMT constraints**

The IoMT devices have restricted computation, bandwidth, and energy. We will explore lightweight cryptography and model compression to reduce overhead and enable wider participation.

- **Balancing privacy–utility trade-offs**

To safeguard patient data, differential privacy is employed, but at the cost of accuracy. Adaptive mechanisms that adaptively control the noise (disorder) are expected to protect the strongest privacy with little harm.

- **Validating with real healthcare data**

Existing benchmarks don't include healthcare-specific traffic or evolving threats. This data will feed experimentation in future research upon tests on hospital and regional datasets to engage zero-day attacks and other dynamic conditions.

10. CONCLUSION

In this study, we introduced health-fusion deep learning, a fusion deep learning framework to detect intrusions on benchmark datasets that are transferable in the healthcare domain. The accuracy performance on the KDD-CUP 99, NSL-KDD, and UNSW-NB15 was > 95% while leakage of privacy on secure aggregation is < 2%, and communication overhead is about 5% when applied to CNN–LSTM models.

Health-fusion deep learning exhibited the best privacy–performance trade-offs against centralized and baseline federated approaches. Although secure aggregation is consistent with healthcare privacy principles, and differential privacy also falls in the general domain of formal compliance, practical deployment support was an open research problem. Important challenges are adversarial robustness, non-IID data, the constraints of the IoMT, and privacy–utility trade-offs.

Federated learning with deep models and privacy-preserving methods leads us to an exciting path that could be scalable for intrusion detection. Further enhancements and testing on real healthcare datasets could make health-fusion deep learning a feasible approach for protecting sensitive healthcare data from emerging cyber-threats.

REFERENCES

[1] Sharma, V., Kumar, M. (2025). Improving intrusion detection with hybrid deep learning models: A study on CIC IDS2017, UNSW NB15, and KDD CUP 99. Journal of Information Systems Engineering and Management, 10(11s): 634-650. https://doi.org/10.52783/jisem.v10i11s.1665

[2] Yao, R.Z., Wang, N., Liu, Z.H., Chen, P., Sheng, X.J. (2021). Intrusion detection system in the advanced metering infrastructure: A cross-layer feature-fusion CNN-LSTM-based approach. Sensors, 21(2): 626. https://doi.org/10.3390/s21020626

[3] Abdallah, M., Le Khac, N.A., Jahromi, H., Jurcut, A.D. (2021). A hybrid CNN LSTM based approach for anomaly detection systems in SDNs. In Proceedings of the 16th International Conference on Availability, Reliability and Security, Vienna, Austria, pp. 1-7. https://doi.org/10.1145/3465481.3469190

[4] Shaikh, J.A., Wang, C., Sima, M.W.U., Arshad, M., Owais, M., Hassan, D.S.M., Alkanhel, R., Muthanna, M.S.A. (2025). A deep reinforcement learning based robust intrusion detection system for securing IoMT healthcare networks. Frontiers in Medicine, 12: 1524286. https://doi.org/10.3389/fmed.2025.1524286

[5] Abbas, S.R., Abbas, Z., Zahir, A., Lee, S.W. (2024). Federated learning in smart healthcare: A comprehensive

review on privacy, security, and predictive analytics with IoT integration. Healthcare, 12(24): 2587. https://doi.org/10.3390/healthcare12242587

[6] Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A. (2009). A detailed analysis of the KDD CUP 99 data set. In 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, Canada, pp. 1-6. https://doi.org/10.1109/CISDA.2009.5356528

[7] Revathi, S., Malathi, A. (2013). A detailed analysis on NSL KDD dataset using various machine learning techniques for intrusion detection. International Journal of Engineering Research & Technology, 2(12): 1848-1853.

[8] 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), Canberra, ACT, Australia, pp. 1-6. https://doi.org/10.1109/MilCIS.2015.7348942

[9] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A., Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, Dallas, Texas, USA, pp. 1175-1191. https://doi.org/10.1145/3133956.3133982

[10] Dwork, C. (2006). Differential privacy. In Automata, Languages and Programming, pp. 1-12. https://doi.org/10.1007/11787006_1

[11] Xu, J., Glicksberg, B.S., Su, C., Walker, P., Bian, J., Wang, F. (2021). Federated learning for healthcare informatics. Journal of Healthcare Informatics Research, 5: 1-19. https://doi.org/10.1007/s41666-020-00082-4

[12] Rieke, N., Hancox, J., Li, W.Q., Milletari, F., et al. (2020). The future of digital health with federated learning. npj Digital Medicine, 3: 119. https://doi.org/10.1038/s41746-020-00323-1

[13] Reyaz, M.A.T., Vanitha, V., Rajathi, N. (2025). Federated learning based intrusion detection system for healthcare domain. In Intelligent Solutions for Smart Adaptation in Digital Era, pp. 117-129. https://doi.org/10.1007/978-981-97-8193-5_11

[14] Shokri, R., Stronati, M., Song, C., Shmatikov, V. (2017). Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, pp. 3-18. https://doi.org/10.1109/SP.2017.41

[15] Issa, M., Aljanabi, M., Muhialdeen, H. (2024). Systematic literature review on intrusion detection systems: Research trends, algorithms, methods, datasets, and limitations. Journal of Intelligent Systems, 33(1): 20230248. https://doi.org/10.1515/jisys-2023-0248

[16] Kimanzi, R., Kimanga, P., Cherori, D., Gikunda, P.K. (2024). Deep learning algorithms used in intrusion detection systems: A review. arXiv preprint arXiv:2402.17020. https://doi.org/10.48550/arXiv.2402.17020

[17] Al, S., Sağiroğlu, Ş. (2024). A review of explainable artificial intelligence in intrusion detection systems. In 2024 17th International Conference on Information Security and Cryptology (ISCTürkiye), Ankara, Türkiye, pp. 1-6. https://doi.org/10.1109/ISCTrkiye64784.2024.10779325

[18] Ali, M.S., Ahsan, M.M., Tasnim, L., Afrin, S., et al. (2024). Federated learning in healthcare: Model misconducts, security, challenges, applications, and future research directions - A systematic review. arXiv preprint arXiv:2405.13832. https://doi.org/10.48550/arXiv.2405.13832

[19] Li, J., Sun, A.X., Han, J.L., Li, C.L. (2022). A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1): 50-70. https://doi.org/10.1109/TKDE.2020.2981314

[20] Dhakal, R., Raza, W., Tummala, V., Niure Kandel, L. (2024). Enhancing intrusion detection in IoT networks through federated learning. IEEE Access, 12: 167168-167182. https://doi.org/10.1109/ACCESS.2024.3495702

[21] Yang, Q., Liu, Y., Chen, T.J., Tong, Y.X. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2): 1-19. https://doi.org/10.1145/3298981