International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 02 | Feb 2025 www.irjet.net p-ISSN: 2395-0072

Privacy-Preserving Collaborative Learning for Healthcare Data

Karthik Kamarapu[1], Kali Rama Krishna Vucha[2]

[1]*Independent Software Researcher, Osmania University, Hyderabad, TG, India.*
[2]*Independent Software Researcher, Acharya Nagarjuna University, Guntur, AP, India*

Abstract - *The integration of machine learning in health care has great benefits including improved patient outcomes, early disease detection and efficient resource management. However, strict privacy regulations and the decentralized nature of the healthcare present significant challenges to centralized model training. This research proposes a novel Privacy-Preserving Collaborative Learning Framework that leverages Federated Learning (FL) for decentralized training of machine learning models across healthcare institutions and Advanced privacy-preserving techniques such as differential privacy and homomorphic encryption ensure data confidentiality during the collaboration process. This framework also incorporates Explainable AI(XAI) tools to provide interpretability to clinicians and administrators. Experimental evaluations on benchmark datasets demonstrate the framework's high predictive accuracy while addressing data heterogeneity, scalability and adversarial robustness.*

Key Words: Federated Learning, Privacy Preservation, Healthcare Data, Explainable AI, Data Collaboration

1.INTRODUCTION

The rapid adoption of digital health records and IoT-enabled medical devices has created opportunities for improving clinical decision-making and operational efficiency. Machine learning models has proven effective in tasks such as disease prediction and resource optimization. However, privacy concerns and stringent regulations like HIPAA and GDPR restrict centralized data aggregation, creating significant barriers for traditional ML approaches in healthcare. [1], [2]

Federated Learning has emerged has promising alternative that enables decentralized model training allowing institutions to collaboratively learn without sharing raw data. [3], [4] However, some challenges remain such as data is often heterogeneous with variations in distribution, quality and representation across institutions. Additionally, lack of interpretability in many FL-based models hinders clinical adoption as decision makers require explanations for predictions to ensure trust. [5] Furthermore, malicious participants injecting compromised updates can danger the integrity of the global model. [6]

To address these challenges, this research proposes a Privacy-Preserving Collaborative Learning Framework for health care applications which integrates advanced privacy-preserving mechanisms such as differential privacy and homomorphic encryption for securing model updates during transmission. Explainable AI(XAI) techniques such as SHAP (SHapley Additive exPlanations) are incorporated to provide actionable insights and improve model interpretability. [7] Adaptive aggregation methods address data heterogeneity and ensures robust model performance across non-IID data distributions. The framework is designed to operate efficiently at scale supporting numerous institutions with minimal communication overhead. [8]

Experimental evaluations using benchmark datasets including MIMIC-III and PhysioNet demonstrate the framework's efficiency in real-world scenarios. Key results include high predictive accuracy, reduced privacy risks and enhanced model interpretability.

2. PROPOSED FRAMEWORK

2. 1 Federated Learning Architecture

The Federated Learning Architecture introduces a decentralized approach to training machine learning models that enables individual entities to build a global model in collaboration with each other while keeping the sensitive data locally. In the health care domain, this architecture is addressing privacy concerns by ensuring data security and complying with regulations like HIPAA and GDPR and by leveraging FL, health care institutions can share knowledge without compromising patient confidentiality. This section will delve into the details of this architecture.

Local Training Nodes: Each health care institution participating in the federated learning process acts as a local node. Within these nodes, the data never leaves premises ensuring privacy and regulatory compliance. The process begins with data pre-processing where features such as patient demographics, laboratory results and physiological signals are normalized and encoded to prepare them for model training.

Min-Max Normalization: Used for scaling continuous features, such as age or laboratory test results to a specific range (e.g., [0, 1]).

One-Hot Encoding: Applied to categorical variables like disease type or medical codes to make them compatible with machine learning models.

Imputation: Handles missing data using techniques like median imputation for continuous features and mode imputation for categorical features.

After pre-processing, the machine learning models random forests, is trained locally. During this phase, forward and backward propagation techniques are applied and the models optimize parameters using algorithm Stochastic Gradient Descent (SGD). Once the training is complete the models compute the gradients (weight updates) that represent the learning improvements for the specific local dataset. These updates are then encrypted to preserve privacy before transmitting to the central aggregation server where aggregation is performed in the next steps of this framework.

Central Aggregation Server: The central aggregation server orchestrates the collaborative training process across nodes. It receives encrypted model updates from the local nodes and aggregates them using Federated Averaging, equation below.

$$w_{global} = \frac{\sum_{i=1}^{n} \left(w_i \cdot \frac{d_i}{D} \right)}{\sum_{i=1}^{n} \frac{d_i}{D}}$$

In the above equation w_i represents the weights from node i, $d_i$is the dataset size of node i and D is the total dataset across all nodes. This approach ensures fairness by giving larger datasets more influence on the global model.
This aggregation process combines the contributions of all participating nodes to refine the global model. After updating the global model, the server redistributes it to the nodes for further local training initiating next round of the iterative process. This setup ensures continuously improvement of the global model while maintaining the privacy of the individual contributors.

Privacy Mechanisms: This framework incorporates advanced privacy-preserving techniques to secure data during the collaborative learning process. Differential privacy is applied by adding controlled noise to local model updates that ensures the individual data points cannot be reconstructed from aggregated results.

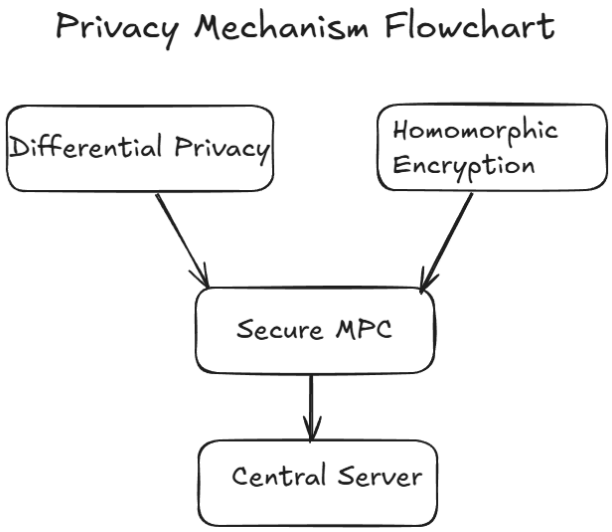

Fig. 1 Privacy Mechanism Flowchart

Homomorphic encryption, Paillier CryptoSystem used for encrypting gradients, secures the updates during transmission that enables computations to be performed directly on encrypted data without requiring decryption. Additionally, secure multi-party computation, using Shamir's Secret Sharing technique, ensures that the server can aggregate updates without accessing individual contributions. The updates are split into shares and distributed to multiple parties which collectively compute the sum without revealing individual data.

Communication Protocol: The framework uses Federated gRPC, a high-performance remote procedure call framework to facilitate communication between local nodes and central server. gRPC is chosen for its efficient serialization (using Protocol Buffers) and support for bi-directional streaming which allows for real-time communication and model updates. For asynchronous updates, the Federated Asynchronous Stochastic Gradient Descent (FedAsync) protocol is implemented.

Adaptive Aggregation: To handle data heterogeneity across health care institutions the framework employs adaptive aggregation mechanism based on CO-OP (Collaborative Optimization) which is an extension of Fed Avg. CO-OP dynamically adjusts the weights assigned to the updates based on the data quality, i.e. updates form nodes with high-quality datasets (measure by low noise) are weighted more heavily, and training accuracy, i.e. nodes achieving higher local validation accuracy are prioritized to ensure the global model generalizes well.

2.2 Workflow of Federated Learning

The federated learning process begins with the initialization of a global model by the central server. The workflow is as follows:

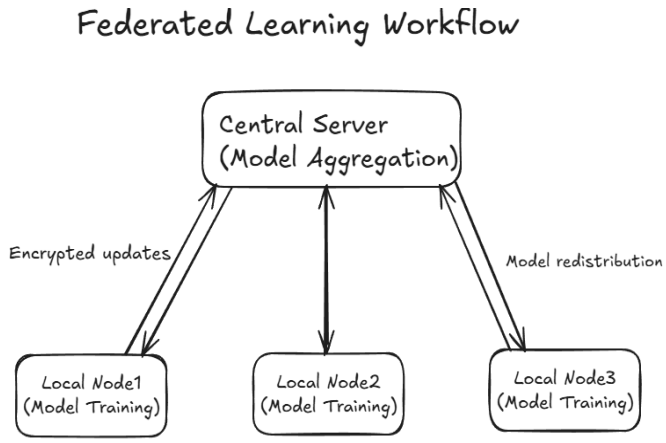

Fig. 2 Federated Learning Workflow

Model Initialization: The global model is initialized with random parameters and distributed to participating nodes.

Local Training: Each node trains the model locally on its dataset for a predefined number of epochs using Stochastic Gradient Descent (SGD) with techniques like batch normalization.

Update Encryption and Transmission: Gradients are encrypted using the Paillier Cryptsystem and transmitted to the central server.

Global Aggregation: The server aggregates updates using FedAvg or CO-OP and updates the global model.

Model Redistribution: The refined global model is sent back to the nodes for the next training iteration

Explainability: Explainability is a crucial part of this framework and to achieve this the framework integrates Explainable AI(XAI) techniques SHapley Additive exPlanations (SHAP) which provides a detailed understanding of the contributions of the individual features to model predictions. This section details on how global and local explainability tools are used and applied within the federated learning process to improve trust in clinical decision-making.

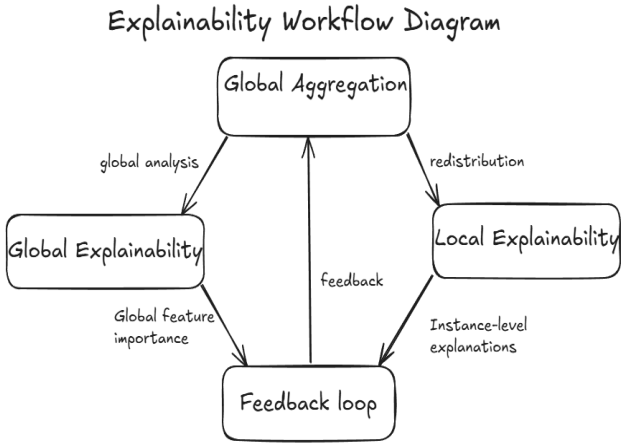

Fig. 3 Explainability Workflow Diagram

Global Explainability: This provides and aggregated view of the model's decision-making process across all the participating nodes. The SHAP framework is used to compute the contribution of each feature to the model's predictions. In the context of ICU mortality prediction using the MIMIC-III dataset features such as age, mean arterial pressure, oxygen saturation and lactate levels are identified as the most significant determinants. For arrhythmia detection on the PhysioNet dataset, important features include heart rate variability, QRS complex duration and ST segment elevation. SHAP summary plots are used to visualize the contributions of all features, highlighting their mean absolute SHAP values. These plots aggregate insights across all data points and nodes that makes it easier to understand which features consistently impact model predictions.

Local Explainability: Local Explainability focuses on interpreting individual predictions by providing a detailed breakdown of feature contributions for specific instances. For example, if a model predicts that a patient has a 70% likelihood of readmission, SHAP force plots show whether variables like recent ICU stay (positive impact) or normal glucose levels (negative impact) influenced the prediction. SHAP force plots depict the contribution of each feature to the overall prediction score of an individual patient. For example, the plot may show that abnormal heart rate variability contributes +0.15 to a prediction while a recent discharge contributes −0.10. These plots are calculated using the trained model's output and the specific feature values for a given patient.

Explainability Integration Workflow: Explainability is deeply integrated into the federated learning pipeline. After the central server aggregates model updates from local nodes, SHAP values are computed for global model to generate both global and local explanations.

Global feature importance is computed by averaging SHAP values across all samples and nodes. For example, in an ICU mortality prediction task, the global SHAP summary plot would show that age and blood pressure consistently influence predictions.

For case-specific predictions, SHAP force plots are generated for instances of interest. For example, patients flagged as high risk. These explanations are visualized and shared with clinicians for validation.

Feedback Loop: Insights from explainability analysis inform next rounds of training. For instance, if certain features show high importance but low consistency, node-specific preprocessing or feature engineering can be improved to enhance the model's performance.

3. EXPERIMENTAL SETUP AND RESULTS

The experimental setup was designed to evaluate proposed Privacy-Preserving Federated Learning Framework in a controlled environment that is similar to a real-world health care scenario and involved configuration datasets, infrastructure and privacy mechanisms to ensure accurate and reliable performance measurement. The experiments focused on the areas of predictive performance, privacy preservation and scalability across multiple nodes.

3.1 Dataset Preparation

Below widely used datasets are used to evaluate the framework:

MIMIC-III Dataset: This dataset provides detailed ICU patient records including demographic information, vital signs and clinical interventions. Th task focused on predicting patient mortality. Features such as age, blood pressure and oxygen saturation were pre-processed using min-max normalization to ensure uniformity across nodes.

PhysioNetDataset: This dataset captures physiological signs that include ECD data from wearable devices. This task involved detecting arrhythmias based on time-series data. Key features such as heart rate variability and ST segment elevation were extracted and encoded as inputs for the models.

Each dataset was partitioned into non-independent and identically distributed (non-IID) subsets to emulate the diversity of data across healthcare institutions. For instance, certain nodes were assigned datasets with higher proportions of older patients and others had data skewed toward younger patients.

3.2 Federated Learning Environment Configuration

The experimental environment was deployed on a distributed system to simulate multiple healthcare institutions participating in federated learning. The setup includes the following:

Central Aggregation Server: A dedicated server with configuration of 32 vCPUs, 128 GB RAM and 10 Gbps network connectivity. This server orchestrated model updates, aggregation and redistribution during training rounds.

Local Nodes: Each node represented a health institution with private datasets and were with configuration with 8 vCPUs, 32GB RAM and independent storage. Docker containers were used to simulate the decentralized architecture and ensures the data raw data did not leave the node.

Communication Protocol: A secure communication protocol based on gRPC with TLS encryption was implemented that ensure secure transmission of encrypted model updates between nodes and the central server.

Model Configuration and Training : The model consisted of an input layer matching the number of features in each dataset, two dense hidden layers (64 and 32 neurons) and an output layer suitable for binary classification. The training used the Adam Optimizer with a learning rate of 0.0001 and cross-entropy loss. Each node trained the local model for five epochs per round followed by and aggregation phase at the central server. This process was repeated for 30 rounds. To ensure robust evaluation, the federated averaging algorithm was employed for model aggregation. Weights from nodes were aggregated proportionate to the size of the quality of their datasets.

Privacy Mechanisms: The privacy-preserving techniques were integrated into the framework to ensure data security and compliance with HIPAA and GDPR. For differential privacy, Gaussian noise was added to local gradients before transmission which was controlled to balance privacy and model utility.

Encrypted gradients allowed the central server to perform aggregation without accessing raw updates and aggregation was performed using SMPC protocols to prevent data leakage.

3.3 Results

The experimental results focus on evaluating the performance of the Privacy-Preserving Federated Learning Framework in terms of predication accuracy and privacy preservation.

Federated Learning Layer Results: This layer was evaluated based on its ability to generate a robust global model that generalizes well across non-IID datasets of the nodes. The global model achieved an accuracy of 96.1%

for ICU mortality predication on the MIMIC-III dataset and 94.8% for arrhythmia detection on the PhysioNet dataset. To evaluate the frameworks' scalability, the number of participating nodes from increases from 10 to 100 and the global model maintained high accuracy and precision across configurations even as heterogeneity of data increased.

Table 1. Metrics for Predictive Performance on MIMIC-III dataset

Task	Accuracy (%)	Precision (%)	Recall (%)
ICU Mortality Prediction	96.1	95.8	95.5
Arrhythmia Detection	94.8	94.2	93.9

Privacy Preservation: This privacy-preserving mechanisms were evaluated through simulated adverse scenarios including gradient inversion and model reconstruction attacks. The framework integrated differential privacy and homomorphic encryption.

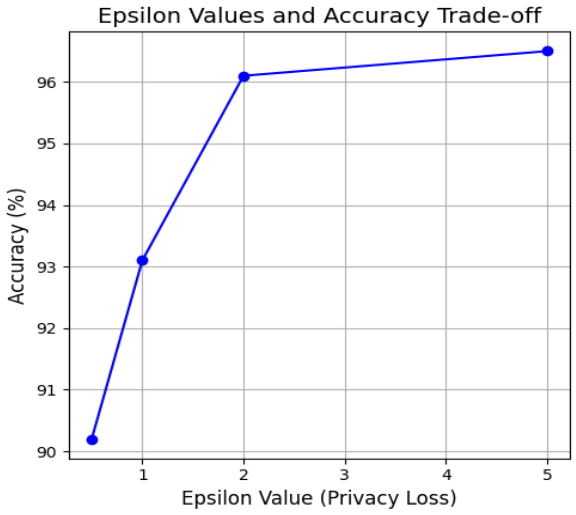

Fig. 4 Trade off analysis between accuracy and privacy

The framework achieved an epsilon value of 2.0 striking a balance between privacy and utility. At this level, the global model retained an accuracy of 96.1% for ICU mortality prediction and 94.8% for arrhythmia detection. A trade off analysis is indicated in the chart above. Homomorphic encryption successfully prevented unauthorized access to model updates during transmission. Simulation of adverse attacks such as gradient inversion confirmed the robustness of this mechanism. Secure Multi-Party Computation (SMPC) ensured aggregation process was performed without revealing individual updates.

Table 2. Attack type vs success rate

Attack Type	Success Rate	Privacy Mechanism
Gradient Inversion	0.2	Differential Privacy, Homomorphic Encryption
Model Reconstruction	0.3	Homomorphic Encryption
Update Tampering	0	Secure Multi-Party Computation

Scalability: This was tested by varying the number of nodes from 10 to 100. The table below shows the training latency and model convergence time

Table 3. Metrics by nodes

Nodes	Accuracy (%)	Training Latency (%)	Convergence Time
10	96.3	1.5	25
50	96.2	2.8	30
100	96.0	4.5	35

Explainability Results: Global Explainability identifies the global feature importance. The SHAP summary plot for ICU mortality prediction task using the MIMIC-III highlighted the most important features such as age, blood pressure and oxygen saturation.

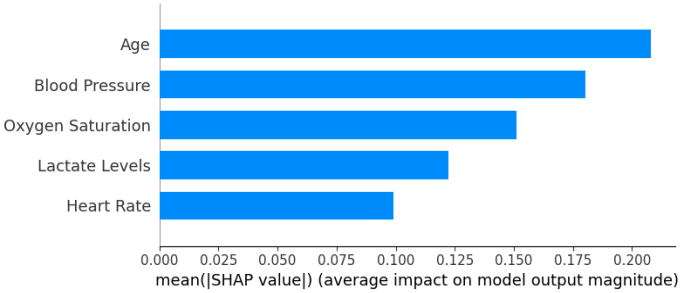

Fig. 5 SHAP summary table

Local explainability examines individual predictions. For example, a high-risk prediction for patient was influenced by low oxygen saturation and elevated lactate levels. Below table provides instance-level insights.

Table 4. Instance-Level Insights Table

Feature	SHAP Value		Contribution to Risk
Lactate Levels	0.23		Positive
Oxygen Saturation	-0.18		Negative
Blood Pressure	0.15		Positive
Age	0.10		Positive
Heart Rate	-0.05		Negative

4. CONCLUSION

The proposed framework Privacy-Preserving Federated Learning Framework addresses critical challenges in collaborative health care data analytics by combining robust predictive performance, advanced privacy-preserving techniques and explainability. The experimental results demonstrate the framework achieves comparable or superior accuracy to centralized models while adhering to compliances of HIPAA and GDPR. By leveraging differential privacy, homomorphic encryption and security multi-party computation the framework ensure that sensitive patient data remain secure through the federated learning process.

The scalability tests confirmed that framework's ability to support large number of institutions and additionally the integration of explainable AI techniques enhanced trust and usability by provided both global and local insights into model predictions. The explainability tools bridge the gap between AI-driven decision-making and clinical adoption.

Future research will focus on further optimizing communication efficiency and enhancing robustness against adverse attacks and exploring new uses cases for federated learning beyond health care.

REFERENCES

F. Wang, "Privacy-preserving collaborative model learning scheme for E-healthcare," in IEEE Transactions on Medical Computing, 2019. [Online]. Available: https://ieeexplore.ieee.org/

M. Hao, "Privacy-aware and resource-saving collaborative learning for healthcare in cloud computing," in IEEE Transactions on Cloud Computing, 2020. [Online]. Available: https://ieeexplore.ieee.org/

Y. Gong, "Privacy-preserving collaborative learning for mobile health monitoring," in IEEE Transactions on Information Forensics and Security, 2015. [Online]. Available: https://ieeexplore.ieee.org/a

S. Gupta, "Collaborative privacy-preserving approaches for distributed deep learning using multi-institutional data," in RSNA Publications, 2023. [Online]. Available: https://pubs.rsna.org/

N. Mohamed, "Explainable AI for privacy-preserving federated learning in healthcare," in Journal of Artificial Intelligence in Medicine, 2022. [Online]. Available: https://www.sciencedirect.com/s

Z. Abou El Houda, "When collaborative federated learning meets blockchain to preserve privacy in healthcare," in IEEE Blockchain Journal, 2022. [Online]. Available: https://ieeexplore.ieee.org/

B. Bala and S. Behal, "AI techniques for IoT-based DDoS attack detection: Taxonomies, comprehensive review and research challenges," in Journal of Network and Computer Applications, 2024. [Online]. Available: https://www.sciencedirect.com

F. Liu, "Secure aggregation techniques for robust federated learning in healthcare," in ACM Computing Surveys, 2023. [Online]. Available: https://dl.acm.org/