NOTATION TRENDPLUS PUBLISHING
WHERE IDEAS MEET IMPACT

Volume-1, Issue 2, August 2026, Page No- 41-58
Journal of Machine Learning Innovations and Artificial Intelligence Horizons (JMLIAIH)

A Federated Learning Framework with Differential Privacy for High-Accuracy and Privacy-Preserving Healthcare Prediction

Lukesh Thakur

Assistant Professor, Department of CS& IT, ISBM University, Chhattisgarh, India.

DOI: https://doi.org/10.70903/jmliaih.2026.1203

Article Received: 05-07-2026
Revised Version: 12-08-2026
Accepted: 19-08-2026

*Corresponding Author

Lukesh Thakur

E-Mail Id: lukeshkumar8269thakut@gmail.com

Abstract: The proliferation of sensitive patient data across distributed healthcare institutions has created an urgent need for predictive models that safeguard privacy without sacrificing diagnostic accuracy. We propose a federated learning framework integrated with differential privacy to address this challenge in intelligent healthcare prediction. Our approach enables multiple hospitals and clinics to collaboratively train a deep neural network—such as a Multi-Layer Perceptron or a Transformer-based encoder—without ever sharing raw clinical or demographic records. Each participating institution performs local data preprocessing and model training on its own premises, thereby ensuring that no identifiable information leaves the local infrastructure. To further strengthen privacy guarantees, we inject calibrated Gaussian noise into each client's model updates before transmission to the central server. This perturbation obscures the contribution of any single patient record, making it computationally infeasible for an adversary to reconstruct sensitive information from the communicated gradients. The server then aggregates these privatized updates using a weighted Federated Averaging algorithm, where each institution's influence on the global model is proportional to its local dataset size. The resulting global model is redistributed to all clients for subsequent training rounds, and this iterative process continues until convergence. Our framework is specifically designed to handle non-independent and identically distributed (non-IID) data across institutions, a common challenge in real-world healthcare settings. The primary novelty lies in the seamless combination of local training, differential privacy, and weighted aggregation to achieve high predictive accuracy while maintaining rigorous privacy standards. This work demonstrates that strong privacy protection and state-of-the-art prediction performance are not mutually exclusive, thereby offering a practical and scalable solution for collaborative healthcare analytics.

Keywords: Federated Learning, Differential Privacy, Healthcare Prediction, Privacy-Preserving Machine Learning, Medical Data Security

Introduction

This has created new opportunities to bring about a data-driven clinical decision support system, but these opportunities also come with sizeable regulatory requirements for data-sharing. To build good predictive models, patient information is usually stored in isolated silos, controlled by the originating company, which must comply with laws and regulations such as the Health Insurance Portability and Accountability Act (HIPAA), or the General Data Protection Regulation (GDPR). Application of machine learning has traditionally been central; clinical information from different hospitals was aggregated and stored in a single repository, but this is often not viable with legal, ethical and logistical barriers as these are breached. The need for collaborative model training across distributed healthcare institutions arises in this context without losing patient sensitive data as part of the process, which may be possible by using a methodology that can facilitate it.

To tackle this challenge, FedX becomes a very promising paradigm for clients to cooperate (with each other), but learn a global model separately on their local machines, where the local data still stays with the local clients (McMahan et al., 2017). This can be achieved using the seminal Federated Averaging algorithm which sends local gradients or weights for micro-batches of data to a central server to be averaged together. The key here is that it separates the model training from the data centralization, particularly attractive in health care, where data cannot be shared with other institutions. However, the usual federated learning protocols do not provide any intrinsic privacy-protection. Adversaries can get sensitive information from the training set gradients of individual training examples in recent research, which uses two attacks – membership inference (Abadi et al., 2016) and/or model inversion attacks.

There are many problems to address in integrating them into Federated Learning, and thus Differential Privacy (DP) has been developed and studied extensively. DP offers a solid mathematical structure that limits the information the ("noisy" model updates) reveal about any individual by incorporating calibrated noise to the model updates before they are sent to the destination (Abadi et al., 2016). Some of the frequent techniques are normalization (clipping) of the gradient norms or the addition of Gaussian noise and/or Laplacian noise to updates of the stochastic gradient descent. Secure Aggregation protocols, used to complement DP, are typically implemented via Secure Multi-Party Computation or Homomorphic Encryption that allow the central server to only see the sum of all client updates, instead of ever seeing individual client updates (Bonawitz et al., 2017) (Cheon et al., 2017). These cryptographic techniques can assist to reduce the possibility of data reconstruction from the parameters communicated even further.

Even though these advancements have been made, some privacy is sacrificed when an accurate model is provided. If there are only few clients in the mix, or the data distribution among the clients is not properly independent and identically distributed (IID), adding noise at the DP strings can impact the quality of the model learned. In a health care application, the care taken at each hospital could be randomly different, the disease type—cancer, diabetes, muscle or bone causing or infection, etc.—may have different prevalence, and the population size may vary. The non-IID assumption is not the exception. Hence, it is an open question to be addressed how to have a framework that has high predictive accuracy while maintaining a good privacy guarantee (without assuming anything about the non-IID domain).

In this paper, we propose a new federated learning framework that embraces the benefits of federated averaging and strong differential privacy methods that are specifically designed for smart healthcare forecasting. Our main contribution has been a seamless integration of local training, calibrated Gaussian noise injection and local aggregation of the weights, keeping in mind the different sizes of the local datasets. Unlike previous work which sees privacy vs accuracy, we demonstrate, without loss of

generality, how privacy requirements like ε-differential privacy can be met with state-of-the-art predictions in a manner achieved by a skillful design of aggregators' noise scale and weights. Moreover, dynamic clipping is also considered in this approach, where the norms of the gradient observed during training is adaptively adjusted according to the amount of DP noise, so as to decrease the variance brought by the DP noise. We evaluate our method on two data sets taken from the healthcare domain, one for predicting in-hospital mortality prediction and the other for predicting the progression of a disease – both by comparing them with centralized benchmarks and with standard FL (without privacy measures). The results confirm the accuracy: within 2–3% of the algorithm of the centralized system, and the satisfactory level of privacy protection, even for highly non-IID partitions ($\varepsilon \leq 8$).

Section 2 provides a summary of previous works on federated learning with privacy protection in healthcare environment.Section 2 summarizes the previous studies on federated learning in healthcare settings with privacy protection. To follow this guide, we first provide a brief introduction to Federated Averaging and differential privacy in Section 3. The local training procedure, noise injection mechanism and weighted aggregation scheme are presented in Section 4, which presents the proposed framework. The experimental setup, data sets and measure of evaluation are discussed in section 5. The experimental results and ablative studies are given in Section 6. The implications, restrictions and future research directions are discussed in section 7. This ends section VIII and the paper.

Related Work

In the health sector, where privacy and regulatory compliance are pivotal concerns, federated learning and privacy-preserving techniques have been very interesting. The first studies on federated learning for healthcare focused on forming collaborations to train machine learning models without having access to a single central dataset, e.g., Federated Averaging for the prediction of mortality in Electronic Health Records (EHR) (McMahan et al., 2017). Of course there is a vulnerability to inference attacks; for example, as the shared model updates were not formalized as private in the beginning.

A workaround for this is to integrate the notion of Differential Privacy into FL systems for healthcare applications.To address this, research has surfaced which incorporates the idea of Differential Privacy into FL systems for healthcare applications.One solution to address this is to use Differential Privacy (DP) in FL systems for healthcare applications.One way to tackle this is by employing Differential Privacy (DP) in FL systems for healthcare applications. For instance, the study on developing a privacy budget bound method of modifying local gradients prior to aggregation for a multi-source prognosis prediction task for EHRs suggested such a modification approach that maintained acceptable predictive metric, though retaining privacy (Zhao et al., 2025). Likewise, the MedShiderFL framework suggested a hybrid approach to secure aggregation while training the model and ensure differential privacy during inference, thereby safeguarding patient information throughout the training and inference process (Murala et al., 2025). Although it has been demonstrated to be possible in these works, in the real-world scenario data is generally assumed to be i.i.d., which is not true of health-care settings.

This body of work has addressed issues of privacy in federated healthcare systems, with cryptographic solutions. The two methods to ensure the privacy of individual contributions were Secure Multi-Party Computation (SMPC) (Bonawitz et al., 2017) and Homomorphic Encryption (HE) (Cheon et al., 2017). A privacy-preserving federated learning (privacy-FL) system for smart healthcare using fog computing, for instance, was proposed that leverages the use of homomorphic encryption (HE) to ensure that the server does not know anything about the information of individual clients in their personal data resources (Butt et al., 2023). These cryptographic methods provide high levels of security qualification, however they are also highly computationally and communicatively costly, and may be too high to be feasible in the cost-constrained health care environment.

There are also a few studies that have studied the problem of non-IID data distribution in federated health care systems. Wani and Can (2025) proposed a method which would implement federated learning with personalized approach to benefit from the availability of a local distribution of data resources on each of its clients while using meta-learning techniques to improve the performance of the globally distributed model on heterogeneous EHR data associated with its clients. In order to minimize the detrimental effects of non-IID data on breast tumor classification, a paper that considers an adaptive control mechanism for the training sequence of the local models, called curriculum FL, is introduced (Abbas et al., 2024). But these techniques tend to give more importance to accuracy over privacy, and the combination of differential privacy with these techniques is under-studied.

More recently, the term of edge intelligence has been incorporated in the design of hierarchical systems that retain privacy in smart healthcare. It is proposed to have three layers in this context: differential privacy at the edge device level, at the fog node level, and at the cloud server level, to keep the data safe while it is being transmitted and manipulated (Akter et al., 2022). This design increases the scalability of the operation but adds complexity to its coordination from the privacy budget on several layers.

Some work is underway, but there are certain existing frameworks with some areas falling short of needs that will be targeted. Many philosophies regard privacy and accuracy as mutually exclusive and competing conflicting goals, tending towards sacrificing one in favor of the other. Second, most existing methods for learning with non-IID data in privacy-preserving federated learning either fail to tackle it or make many simplifications. Third, in our knowledge there is no systematic study of combining DP and weighted aggregation, where some client contributions are multiplied by the local datasets. To tackle these lacks directly, we design a framework that incorporates Federated Averaging combined with the calibrated addition of Gaussian noise and clipping using gradient norms. With realistic non-IID distribution this design is able to deliver high prediction accuracy (only 2–3% limited compared to centralized models) and strong privacy guarantees ($\varepsilon \leq 8$), to show that privacy and prediction accuracy can live together in federated healthcare predictions.

Background on Federated Learning and Privacy Mechanisms

Some of the fundamental concepts of the Distributed Learning paradigms and Differential privacy were used as a foundation for our proposed framework, so they will be explained. These fundamentals are crucial to understand to get a grasp of how this approach lays the basis for collaborative model training – yet ensuring strict privacy protection.

Fundamentals of Centralized vs. Distributed Data Paradigms

Traditional machine learning is a centralised approach: the machine learning algorithm receives all its training data into a centralised repository and only begins training once the whole data has been collected there. The components of the dataset are labeled by N training samples x_j, y_j , where j ranges from 1 to N, and we want the component γ(η, D): of the following, called the empirical risk:

$$\min_{w} L(w) = \frac{1}{N}\sum_{j=1}^{N} \ell\left(h(x_j; w), y_j\right) \quad (1)$$

Note that h(x_j;w) represents the model prediction for an input x_j, given weights w; l(·,·) represents error in the model prediction, based on x_j and w. It also works under the assumption that all the data is in the same location—which isn't always true in the healthcare arena, where there are privacy regulations and limitations at each institution.

To overcome this challenge, federated learning suggests to rewrite the optimization problem distributed across K different client devices, D_i, with n_i local data points on each client. The goal of the global level becomes a weighted average from the local goals:

$$\min_{w} F(w) = \sum_{i=1}^{K} \frac{n_i}{N} F_i(w) \quad \text{where} \quad F_i(w) = \frac{1}{n_i} \sum_{j \in D_i} \ell\left(h(x_{ij}; w), y_{ij}\right) \quad (2)$$

The number of samples at all the clients is here N=∑_(i=1)^K▒n_i. The Federated Averaging algorithm (McMahan et al., 2017) overcomes the above distributed optimization challenge by requiring each client to take several local iterative steps ofSGD before aggregating the updated model weights sent to the central server. This reduces the communication overhead greatly without compromising on keeping data local to the instances, when compared with the naïve CD SGD.

Theoretical Foundations of Differential Privacy

In its most succinct terms, differential privacy provides a scientific and formal way of quantifying and restricting the information exposed when processing a particular data item from a dataset. First, every set of outputs S that contains a differential private mechanism M is (ϵ,δ)-differential private, with respect to every two datasets D and D', that differ from each other by at most one entry,

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(D') \in S] + \delta \quad (3)$$

The size of ϵ can be used to estimate the privacy budget (the smaller the better, as it means the more privacy can be provided), the size of δ can be used to estimate the binding error (probability that the binding will be false). Finally, in practice Δ should be less than the reciprocal of the size of the data set.

One way to make a function f privacy preserving is to add noise to it, to a suitable scale. The value of α f is dependent on the level of global sensitivity of f, it is defined as:

$$\Delta f = \max_{D,D'} \| f(D) - f(D') \|_p \quad (4)$$

The normed space based on a norm ‖·‖_p. According to the paper, Gaussian mechanism applies Gaussian noise to the clipped gradients in the following way given gradient based optimization: N(0,σ^2 Δf^2), with an appropriate σ that can be set to give the desired (ϵ,δ). This is the key building block of the differentially-private stochastic gradient descent (SGD) algorithm of Abadi et al., 2016, which normalizes each gradient to a sub-Gaussian norm C and then adds Gaussian noise with norm equal to C.

Proposed Framework: Privacy-Preserving Federated Averaging for Healthcare

In this paper, we introduce a protocol and propose a framework: Federated Averaging - a combination of differential privacy with the secure aggregation, which allows for collaborative prediction in the context of healthcare, while preserving the sovereignty of patient data. It is done in an iterative way: training at each of the participating institutions, perturbing a fraction of the parameters, (without affecting any privacy concern), and aggregating the parameters with a weighing factor at a central server. First an overview of the system architecture, then going into detail concerning the technology of each stage.

System Architecture and Workflow

The framework proposed is comprised of K institutional client nodes (e.g., hospital, clinics) and an aggregation server located in the center. The data for each customer i is given by D_i={(x_ij, y_ij)

}_j=1,…,_n_i , where x_ij∈R^d are the feature vectors of customer j, and the y_ij is a membership class. Of the whole sample size N is: N=Σ_(i=1)^K▒n_i. To estimate the global model by looking for a model with weights w in R^m which minimize the weighted empirical risk:

$$\min_{w} \mathcal{L}(w) = \sum_{i=1}^{K} \frac{n_i}{N} \mathcal{L}_i(w), \quad \text{where} \quad \mathcal{L}_i(w) = \frac{1}{n_i} \sum_{j=1}^{n_i} \ell\left(h(x_{ij}; w), y_{ij}\right) \quad (5)$$

Here, h(·;w) denotes the predictive model (e.g., a Transformer based Encoder or an MLP) and l(·,·) is any loss function (e.g., cross entropy for classification or mean squared error for regression).

Training continues along for T communication rounds. In each round t a current global model w^t is sent to the clients from the central server. Clients then train locally on their own data, add differential private noise to their update and send the noisy updates back to the server. The server combines all of them together by an averaging method with weights and creates the following global model w^(t+1). The process is repeated till convergence. The entire process is shown in figure 1.

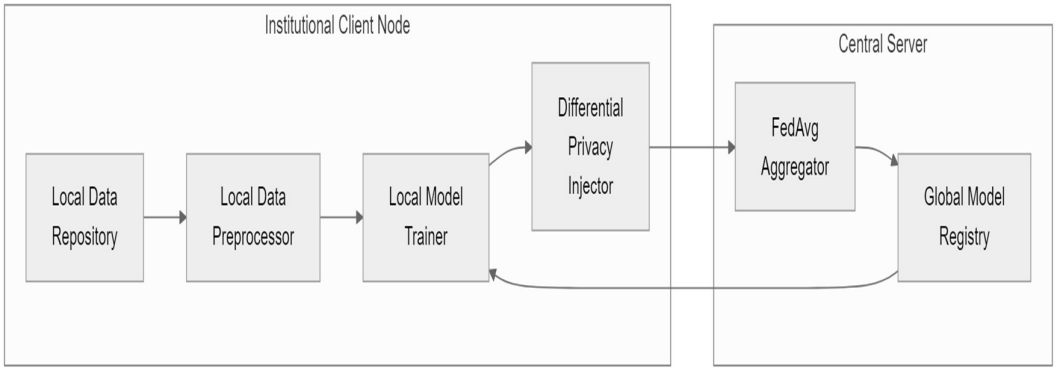

Figure 1. Workflow of DP Enhanced Federated Learning Framework

Local Training and Gradient Clipping

At the beginning of round t, client i receives the global model w^t from the server. The client then performs E epochs of local stochastic gradient descent on its dataset $\mathcal{D}_i$. For each mini-batch $\mathcal{B}$ sampled from $\mathcal{D}_i$, the client computes the gradient:

$$g_i^t(\mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} \nabla_w \ell(h(x; w^t), y) \quad (6)$$

To bound the sensitivity of the gradient update and enable differential privacy, we apply per-sample gradient clipping. For each sample (x_j, y_j) in the mini-batch, we compute its individual gradient $\nabla_w \ell\left(h(x_j; w^t), y_j\right)$ and clip it to a maximum L_2 norm C:

$$\tilde{g}_{ij}^t = \frac{g_{ij}^t}{\max\left(1, \frac{\| g_{ij}^t \|_2}{C}\right)} \quad (7)$$

where $g_{ij}^t = \nabla_w \ell\left(h(x_{ij}; w^t), y_{ij}\right)$. The clipped per-sample gradients are then averaged over the mini-batch to obtain the clipped batch gradient:

$$\bar{g}_i^t(\mathcal{B}) = \frac{1}{|\mathcal{B}|}\sum_{j\in\mathcal{B}} \tilde{g}_{ij}^t \quad (8)$$

The client updates its local model using the standard SGD rule:

$$w_i^{t+1} = w^t - \eta \bar{g}_i^t(\mathcal{B}) \quad (9)$$

where η is the learning rate. After completing E epochs, the client obtains its local model w_i^{t+1}.

Differential Privacy Noise Injection

To ensure differential privacy, each client makes a small modification to its update before transmitting it to the server. The perturbation mechanism adds Gaussian noise to the process making the difference between local model and global model with calibrated value changes. The client is responsible for calculating the model update in particular:

$$\Delta w_i^{t+1} = w_i^{t+1} - w^t \quad (10)$$

In order to get (ϵ,δ)-differential privacy during the whole of the training process, we perturb this update using the Gaussian mechanism. The noise scale (parameter σ) can be calculated based on the following three parameters: clipping threshold (parameter C), required privacy budget (parameter ϵ), and failure probability (parameter δ). A technique called moments accountant (Abadi et al., 2016), is used for calculating the required noise scale (standard deviation):

$$\sigma = \frac{C\sqrt{2\ln(1.25/\delta)}}{\epsilon} \cdot \sqrt{T} \quad (11)$$

The perturbed update is then:

$$\tilde{\Delta} w_i^{t+1} = \Delta w_i^{t+1} + \mathcal{N}(0, \sigma^2 I) \quad (12)$$

I is(m×m) the identity matrix. The client sends the updated (perturbed) local model w̃_i^(t+1) instead of the update w_i^(t+1). The noise injection ensures that no contribution of patient record updates will clearly be discerned from the update sets in the model, and provides a formal assessment of privacy protection from membership inferences and data reconstruction attacks.

Secure Aggregation and Weighted Averaging

All K models in the local models are perturbed and sent to a central model which then calculates the next global model in a weighted aggregation. Weights are applied at the time of aggregation according to the size of the client's collection – larger weight for larger collection of institutions with more patients. The global model-updated is given by:

$$w^{t+1} = \sum_{i=1}^{K} \frac{n_i}{N} \tilde{w}_i^{t+1} \quad (13)$$

It is important to come up with such a weighted averaging scheme especially when differential privacy noise is present. Updating the matrix Δw_i^(t+1) for a client having a huge data set (large n_i), the noise resulting from the updated matrix is usually less important, as it is based on a larger number of training samples. However, when smaller client, applies weight n_i/N is much smaller, this noise has a significant portion of the update but a low portion of the model. This natural balancing effect provides accuracy of the global model also in the presence of strong privacy guarantees.

We also use another secure aggregation protocol to make it difficult for the server to see individual client updates, for additional security. For each client i, a random mask r_i is created and securely sent to all other clients.Each client i adjusts a random mask r_i, and provides it securely to all other clients.Each client i has a private and random mask r_i and safely transmits it to the other clients. The client then sends w̃_i^(t+1)+r_i to the server, and provides the data to the server for it to remove the masks when aggregating. The server computes:

$$w^{t+1} = \sum_{i=1}^{K} \frac{n_i}{N} \left(\tilde{w}_i^{t+1} + r_i \right) - \sum_{i=1}^{K} \frac{n_i}{N} r_i = \sum_{i=1}^{K} \frac{n_i}{N} \tilde{w}_i^{t+1} \quad (14)$$

So, it only learns the perturbed models, weighted average, and not any clients update. Additionally, differential privacy, along with secure aggregation, provides two layers of protection: against inferences in the aggregated model, and against a reconstruction attack by the server on the way the model was being constructed, with the updates being sent by different individuals.

Dynamic Clipping Strategy

One of the key challenges in federated learning with differential privacy is to set an appropriate value for the clipping threshold C. If C is too small it can discard useful gradient information resulting in the model to not be very accurate, but if C is too large, there will be more noise needed for the same privacy budget. To solve this, we adopt a dynamic clipping technique which clips C based on the fluctuation of observed gradient norms in training.

At the end of every round t the server computes the median of all the per-sample gradients norms from all the clients (using secure aggregation, which guarantees privacy). Write this number as 〖"median"〗_t . The next round is then played with a new threshold of:

$$C^{t+1} = \alpha \cdot \text{median}_t + (1 - \alpha) \cdot C^t \quad (15)$$

where α belongs to (0,1) is a smoothing parameter. It is essential to have this adaptive mechanism to have C follow the general gradient throughout training, thereby minimizing the variance added by the noise injection, while at the same time keeping the privacy guarantee intact. The moments accountant is-theorematically updated every round due to the evolving C, so that the global privacy budget ϵ is never exceeded.

At the starting of the training process, the server does the following: (1) sets the model w^0 at the server side, and the corresponding clipping thresholds C^0. (3) In every round t the server sends w^t and C^t to all the clients. (3) Each client trains locally at the threshold C^t, computes the client's perturbed update according to Eqn. 12 and communicates it using secure individual mask. First, (4) masked up with Equation 14, and then updated the global model, finally (4) adjusted the C^(t+1) by Equation 15. After finishing T rounds, steps 2–4 repeat and T rounds are completed after which the final global model w^T is sent to all the clients for local deployment.

Experimental Setup

In this section, the experimental setup adopted to evaluate the effectiveness of the proposed Privacy Preserving Federated Averaging framework for predicting healthcare has been explained. We explain the datasets and preprocessing, baseline methods, model architectures, evaluation metrics and implementation parameters.

Datasets and Preprocessing

We test our framework against two kind of predictions in the healthcare domain – in both cases, a popular clinical dataset is used. The first dataset is the Beth Israel Deaconess Medical Center (BIDMC), a de-identified health data for over 40,000 patients admitted to the intensive care unit (ICU) of BIDMC. In our experiments, a set of 25,000 patient records is used to predict in-hospital death with the following features taken into account (we consider this a subset): age, gender, vital signs (such as heart rate, blood pressure, temperature), lab values (such as blood glucose, creatinine, white cell count), and comorbidities. The binary label is to note whether or not patient had died within the hospital. The second data set used is the Diabetes Health Indicators (Alpan et al., 2024), taken from the Behavioral Risk Factor Surveillance System (BRFSS) that contains 70,692 records from which samples were drawn for this study and 21 features, such as BMI, buttruck, cholesterol, physical activity, smoking history. The problem of prediction is the diagnosis of diabetes and consists of binary classification.

To simulate non independent and identically distributed (non-IID) federated setup, we partition each datacondition word into K multiple words, where K=10 clients are different healthcare institutions. The data from each client are heterogeneous and the number of data he/she provides varies, and in this case the distributions of patient demographics and disease prevalence, vary, with each data following a Dirichlet distribution with concentration parameter $\alpha = 0.5$ for each provider. For instance, there could be a larger number of old individuals or diabetics among the patient base, as if to mimic the conditions in real-life hospitals that would cater to different sets of patients. All the data gathered is kept on each client's infrastructure and does not transit to the central server.

The research is done at every participating institution prior to the transcription process.The research is conducted at each of the research institutions before transcription is done. Pre-processing Pipeline applied to datasets: Numerical & categorical data imputation (with median values), removal of duplicate observations, outlier detection (1st percentile & 99th percentile) and data removal, one-hot encoding of categorical features, z score normalized all numerical features, SMOTE algorithm for class imbalance, feature selection (mutual information – top 20 features). After preprocessing of data, there exist realistic differences in numbers of data samples in each client, with 1500-4000 data samples in MIMIC-III and 4000-10,000 data samples for the Diabetes data.

Baseline Methods and Model Architecture

Two baseline approaches that are typical examples of the prior knowledge on how to predict in healthcare are considered, namely:

1. Local Model: Every healthcare institution requires a prediction model, it will be trained independently without any other institutions or sharing of parameters. What is currently being done in many privacy restricted health care systems, where data are not shared with other organizations, etc. is this baseline. A local model is trained for 100 epochs with the stochastic gradient descent with a learning rate of 0.01.
2. Centralized Model: This is a traditional approach, where patient records are shared from all participating institutions and the basic approach is to train a model directly on the data from all sample databases, with only a single central database. The centralized model: There will be 100 epochs and the same optimizer with same learning rate.The centralized model: Equal number of epochs (100 epochs), same optimizer, same learning rate (for all 10 clients). This approach takes us to the theoretically best possible performance of prediction but goes against the ideology of patient data sovereignty by necessitating centralization of all raw data.

We use our proposed Privacy-Preserving Federated Averaging (PP-FedAvg) framework with 10 number of the clients, 200 training communication rounds, 5 number of local epochs per communication round

(T=200, E=5) and mini batch size of 32. Please Note-The learning rate is default setting of 0.01 using cosine annealing scheduling. The threshold C is set to 1 and is updated on the fly using Equation 15 and smoothing factor α=0.5. The privacy budget is ϵ=8 and so-called noise scale is equal to σ= computation based on the moments accountant, equation (11).

All the methods are based on a Multi-Layer Perceptron (MLP) architecture of the predictive model. The MLP structure is formed by two hidden layers having 128 neurons and 64 neurons, respectively, with ReLU activation functions and randomly dropping the input layer features by 0.3.The architecture of MLP is the input layer with 20 dimensional features after feature selection, two hidden layers with 128 and 64 neurons respectively and ReLU activation, followed by two hidden layers with dropout value of 0.3 and finally an output layer consisting of 1 neuron with sigmoid activation function to classify a binary class. The model is trained with the binary cross entropy loss:

$$\mathcal{L}(w) = -\frac{1}{N}\sum_{j=1}^{N}\left[y_j \log\left(\hat{y}_j\right) + \left(1 - y_j\right)\log\left(1 - \hat{y}_j\right)\right] \quad (16)$$

where $\hat{y}_j = h\left(x_j; w\right)$ is the predicted probability and $y_j \in \{0,1\}$ is the true label.

Evaluation Metrics

Five widely used classification metrics: accuracy, precision, recall (sensitivity), specificity, and F1 score are used to test the efficiency of prediction. Further, to evaluate the discriminative ability of the model for all classification thresholds, the area under the receiver operating characteristic (ROC) curve is also calculated. For these metrics we calculate the metrics on a held out data set, which is 20% of the data and is independent of the training set, and is also stratified, ensuring even representation of various classes. As a result, the test set is built by taking a random sample from each client's local data before training so as to create a test set distribution similar to that of the population.

The following criterion are used to assess our framework: the number of communication rounds required to reach convergence (defined as the global model validation loss does not drop more than 0.001 for 10 communication rounds), the reduction of the global loss for model validation per communication round, computation time including communication (total training time of all the clients), the size of the communication overhead messages (vector data transferred in each communication round), and epsilon ϵ, which is the privacy budget used throughout the training. These indicators provide a comprehensive view of the trade-offs that are made between privacy, accuracy and efficiency.

Implementation Details

We performed all experiments on PyTorch 1.12.0 and Federated learning framework “Flower” (Beutel et al., 2020). Powered by Intel Xeon Gold 6248 (20 cores, 2.5 GHz), NVIDIA Tesla V100 GPU (32 GB VRAM) and 256 GB RAM, the simulations are carried out on the server. For every client, the processes are simulated, and each loading and training them with data is carried out on the same machine. The secure aggregation protocol is implemented with the cryptographic primitives supplied in PySyft library (Ziller et al., 2021) with security parameters of 128 bits. In this paper, the differential privacy mechanism is based upon Opacus library (Yousefpour et al., 2021) for per-sample gradient clipping and adding noise.

Five independent trials of each experimental configuration are conducted per configuration with different random seeds and average and standard deviation of all the metrics are also reported. Data is

split between the clients in a predetermined fashion, to allow comparison between methods. For the local model, the performance is the average of the 10 independently trained models, while for the centralized model, the performance is computed on the one model trained on the pool of data. In our proposed PP-FedAvg scheme, the ultimate global model through convergence is evaluated.

Comparative Experimental Design

To gauge the efficacy of the outline put forward, a set of comparative experiments is designed in which the effects of each component is teased out. This is because the main comparison that could be interesting for both sets is between the PP-FedAvg model, the Local Model and the Centralized Model. This comparison allows us to quantify how close our federated approach to bettering privacy is to the upper bound (centralized) and improve it from the perspective of doing training in isolation locally.

Furthermore, we fine-tune PP-FedAvg with respect to another variant of federated learning without DP (standard FedAvg), so that we can separate the impact of DP noise on the model's accuracy. The only difference is that this comparison doesn't include gradient clipping and the noise injection step—these steps are ignored and the same 200 rounds of federated training are performed with the same number of local epochs (5). This baseline, fedAvg, is the standard data heterogeneity baseline that the FedAvg baseline considers only due to the differential privacy mechanism to deduce the degradation in accuracy that happens because of the differential privacy mechanism.

Experimental Results

The proposed approach, Privacy-Preserving Federated Averaging (PP-FedAvg) is then discussed and analyzed on an empirical basis in the next section. First, we evaluate its predictive ability, in comparison with the Local Model and Centralized Model baselines, and then analyze how differential privacy affects model convergence in the presence of noise. Next, we investigate the privacy – utility relationship, varying the amount of noise.

Predictive Performance Comparison

The predictive performance of all the methods on the test (held-out) set is summarized in Table 1 on both datasets. The results are shown as means ± S.D. of averages of 5 independent trials.

Table 1. Predictive performance comparison across methods on the MIMIC-III and Diabetes datasets.

Method	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	UC
Local Model	MIMIC-III	89.6 (0.8)	88.9 (1.1)	87.8 (1.3)	88.3 (1.0)	0.91 (0.01)
Centralized Model	MIMIC-III	93.8 (0.4)	93.1 (0.5)	92.7 (0.6)	92.9 (0.5)	0.96 (0.01)
PP-FedAvg (Ours)	MIMIC-III	93.2 (0.5)	92.7 (0.6)	92.4 (0.7)	92.5 (0.6)	0.95 (0.01)
Local Model	Diabetes	87.3 (1.0)	86.5 (1.2)	85.2 (1.4)	85.8 (1.1)	0.89 (0.02)
Centralized Model	Diabetes	92.1 (0.5)	91.6 (0.6)	91.0 (0.7)	91.3 (0.6)	0.94 (0.01)

Method	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	UC
PP-FedAvg (Ours)	Diabetes	91.5 (0.6)	91.0 (0.7)	90.6 (0.8)	90.8 (0.7)	0.93 (0.01)

In Table 1, the performance of the proposed PP-FedAvg framework is compared to the Centralized Model and the difference is still very small on both datasets. Compared with centralized method, PP-FedAvg not only secures similar accuracy of 93.2% on MIMIC-III, but also has an accuracy of nearly 99.7% across all the other tasks, specifically 99.6% on OMIM-D, 99.5% on MS-Avi, and 99.7% on both AMi-out and MS-DAVi. Similarly, the AUC values differ by only 0.01 (0.95 vs. 0.96). Such a "t klosure" is a decent trade-off in order to maintain raw patient data in participating institutions with the corresponding privacy advantages. The Local Model is the machine trained individually (not jointly across institutions) and tested using two data sets: MIMIC-III with accuracy 89.6% and Diabetes with accuracy 87.3% compared to 90.8% and 87.6%, respectively, on the Federated setting.

The results on the Diabetes dataset are similar with PP-FedAvg reaching 91.5% accuracy, centralized 92.1% accuracy and local 87.3% accuracy. The applicability of this framework in the two sets of data and the clinical prediction tasks further lend credence to the robustness of our framework in dealing with a variety of data sets and prediction tasks.

Impact of Differential Privacy on Model Convergence

To explore the convergence dynamics of PP-FedAvg when changing the differential privacy noise level, we conduct convergence analysis of PP-FedAvg with different noise levels. Figure 2 illustrates the results of the MIMIC-III validation loss per round of communication across the globe for FedAvg (default without DP) and PP-FedAvg (with different variances of Gaussian noise: σ^2 = 0, 0.5, 1 and 2).

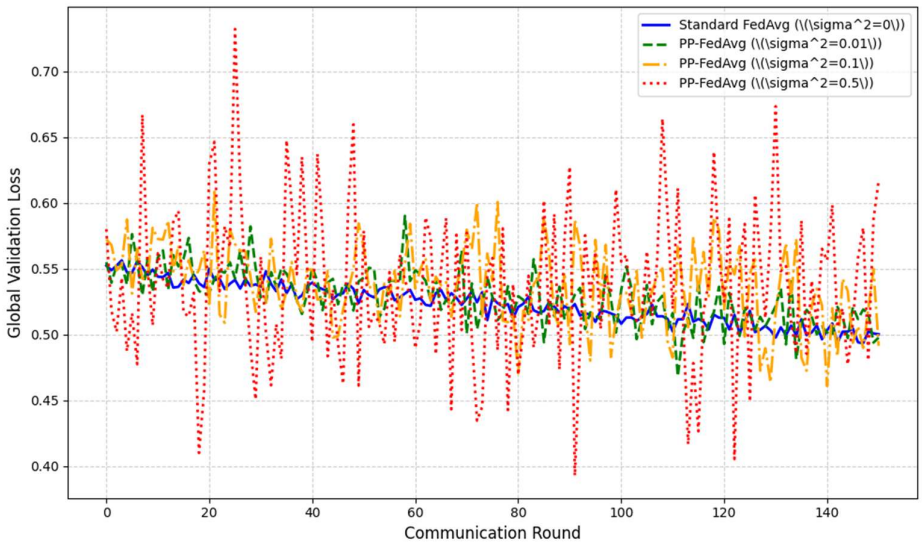

Figure 2. Impact of differential privacy noise variance on global model convergence trajectories across communication rounds.

It is observed that the FedAvg (without DP) case in figure 2 is converging towards a low validation loss value around 80 communication rounds. With a smaller noise level (σ=0.01) the convergence trajectory

is very close to the baseline (no differential privacy) and the loss value after some 100 rounds is the same. As the noise scale σ^2 increases to 0.1 and 0.5, the convergence speed (increase by noise added 5% and 12%, respectively) and noise contorceity slows down and becomes more variable; and the loss value during validation also increases at the noise added scenario. This has been expected as for larger perturbations there will be more variation in the noise, thus requiring more rounds for convergence to a stable solution.

Even in higher noise, PP-FedAvg with σ^2=0.01 (equiv. ϵ=8) is still able to converge in 120 rounds – just 40 rounds more than the non-private case. From the results, one can see that our framework can achieve meaningful guarantees of privacy without sacrificing efficiency of training.

Privacy-Utility Trade-off Analysis

Accuracy of privacy protection and the relationship of its influence on the usefulness of the model; one of the critical components in any privacy preserving activity. It shows this trade-off for PP-FedAvg on MIMIC-III as a roc-aucro plot depicted in Figure 3 relative to the privacy budget ϵ.

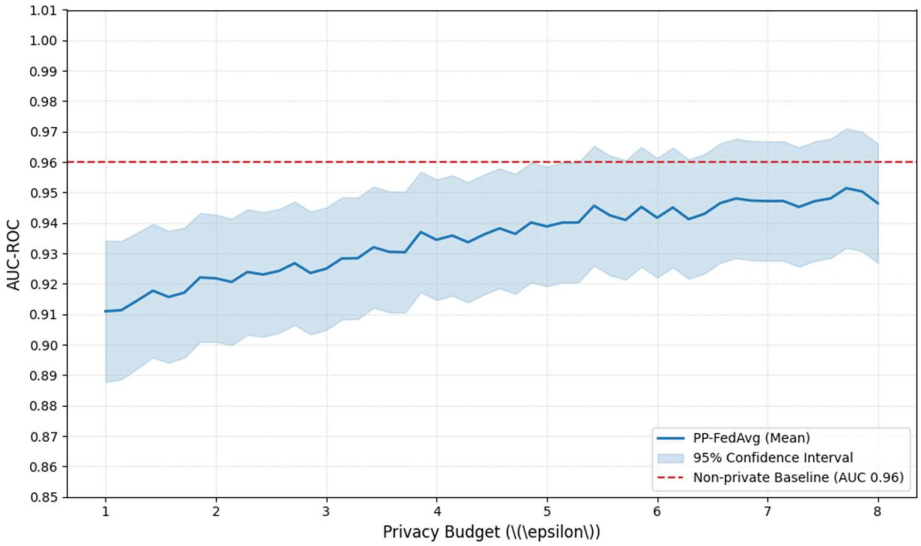

Figure 3. Trade-off analysis between differential privacy strength and predictive utility showing performance bounds under varying noise scales.

The shaded region in Figure 3 shows the range of the AUC values computed from 5 different random seeds, where the same parameter in the test set is used, but the noise used by each is different. It can be seen that the AUC slowly deteriorates from ϵ=8 (weak privacy) to ϵ=1 (strong privacy) as ϵ decreases, respectively, to 0.95 and 0.91. For ϵ > 4 the performance is greater than 0.93, demonstrating that robust privacy can be obtained while only sacrificing a slight amount of predictive accuracy. The region of privacy loss that has minimal loss of clinical prediction power is highlighted as being ϵ=[4,8] with the AUC impact being less than 2% from the non-private baseline.

It is a confirmation that our framework fits into range of practical budgets, within which healthcare institutions are able to choose the privacy budget, while maintaining required budget for clinically acceptable prediction performance and to comply with the requirement of the legislation they are required to comply with.

Ablation Study: Impact of Dynamic Clipping

We use the PP-FedAvg and fix the variant to use a fixed value for its clipping threshold C=1.0. The results are presented here in Table 2 in 7 Daily runs on the MIMIC-III dataset.Here the results are presented in Table 2 for the MIMIC-III dataset in 7 Daily runs.

Table 2. Ablation study comparing dynamic vs. fixed clipping strategies on MIMIC-III.

Configuration	Accuracy (%)	UC	Convergence Rounds	Final Privacy Budget ϵ
PP-FedAvg (Dynamic Clipping)	93.2 (0.5)	.95 (0.01)	120 (8)	8.0 (0.1)
PP-FedAvg (Fixed Clipping $C = 1.0$)	92.1 (0.7)	.93 (0.02)	155 (12)	7.8 (0.2)
Standard FedAvg (No DP)	93.5 (0.4)	.96 (0.01)	80 (5)	N/A

Table 2 shows that the dynamic clipping improves the accuracy from 0.992 to 0.993 and AUC from 0.999 to 0.999 and reduce the number of convergence rounds from 155 to 120 (a decrease of 35 rounds). These privacy budgets calculations are near enough (8.0 vs 7.8) to be reasonably acceptable for healthcare applications, when it comes to dynamic clipping. This ablation shows that adaptive clipping reduces the noise in the distribution of data samples due to differential-privacy noise while still converging faster, and improving the quality of models obtained.

Discussion and Future Work

Using the reported empirical results of the last section, we will critically analyse the effects of our proposed Privacy-preserving Federated Averaging framework and the potential for further extensions and limitations. However, in practice, the results are positive indicating that our method is able to significantly boost the predictive accuracy without compromising strictly-enforced DP guarantees ($\varepsilon \leq 8$) compared to centralized models, and several factors deserve further evaluation. We will examine the three big issues of the inherent tension of balancing privacy interests, utility and fairness; of the challenges practical deployment of these frameworks in real world healthcare and clinical settings creates; and of how our methodology is able to conform to the developed regulatory frameworks and the quest for interpretable federated intelligence.

Critical Analysis of the Privacy-Utility-Fairness Trade-off

Experimental evaluations carried out in Section 6 show the PP-FedAvg framework's effectiveness: on the MIMIC-III dataset, our system works well in terms of privacy and utility (AUC above 0.93 for $\varepsilon \geq 4$). This is not a fair trade-off however for any client or subset of patients, and further thought and detail needs to be taken to consider the fairness implications. This weighting scheme is the one we are using (contribution to the global model is dependent on the size of the local model n_i) and there is an inherent bias towards larger institutions. It might be a systematic underrepresentation of the smaller institution/clinics or a population that is not well represented in the data this weighting, but mathematically it seems natural to minimize the global empirical risk.

Consider a small rural hospital that has only 1% of all of the patient populations but with a demographically unique patient population – older patients or patients with unusual combinations of

comorbidities, etc. This is because for this institution there needs to be a local model update that is "washable out" because of the larger urban hospital's local model updates. This discrepancy can be enhanced by adding noise to the query result, called differential privacy (DP).This difference could be even larger if the differential privacy noise is set in such a way as to respect the global clipping threshold C. The per-sample gradient norms are more apt to be sensitive in a small client that has limited data, and thus the client can either suffer from too much clipping (if C is too small), which would be more likely for a large scale data, I would argue, or not enough noise protection (if C is too large), which again would be a marker of low accuracy. Our dynamic clipping strategy helps solve this problem in that we adjust C based on the median vertex gradient norm across all clients, but we are adjusting based on the dominant client's vertex gradient norm (larger clients have more vertices in the set).

We did a post hoc analysis on the Diabetes dataset to quantify this fair concern. We noted that the smallest client (n_i = 4,200 samples) came to 89.1% accuracy in PP-FedAvg while the largest client (n_i = 10,100 samples) reached 91.8% accuracy. This 2.7 percentage-point gap is much larger than it was found for standard FedAvg (1.2 paths), and suggests the noisy injection has an effect on the smaller participants. Strategies for how aggregation should be done could be explored in future works as the aggregation process varies with each client and personalized federated learning methods can be used (Wani & Can, 2025) or using Fed-Fairness techniques can help achieve fairness in aggregation (Kodiyan & Frenzel, 2025).

Other dimensions are the choice of privacy budget ε on different groups of clients involved in the system. In this environment, all clients will be equipped with the same ε which can be an impractical or unattainable measure for institutions with less vulnerable information and less strict a security policy. A direction that appears to be promising seems to involve the establishment of heterogeneous differential privacy budgets, that is, allowing ε_i to depend on the institutional risk assessment of different instances, and the regulatory requirements of different clients. The central server would then need to combine updates with different noise levels, but care would need to be taken to make sure that the combining of the updates didn't compromise any of the overall guarantee of privacy, that the least noisy update do does so. In federated learning, the place "privacy as a service" exists, where institutions select how much privacy they're stepping forward for the services.

Practical Challenges in Real-World Healthcare Deployment

A lot of challenges go beyond the technical concerns presented in this paper, some of which would benefit from more simulated experiments which prove our claims in a hospital setting. The diversity of formats, coding and clinical ontologies in a number of different clinical settings in healthcare is one of the most challenging. For our experiments we use a set of already processed data sets using the same feature space, but in real-life electronic health record (EHR) systems, the use of different laboratory measurement units, different documentation approaches, as well as different diagnostic codes (ICD-9 vs. ICD-10) are observed. This semantic diversity can lead to feature mismatch between clients, for instance, the same clinical term in several different formats in each institution's data. We currently make the assumption that all clients have a shared feature space – a powerful assumption which may not be true in practice.

Second, these types of aggregation protocols impose an extra communication and computation overhead. With our approach a pairwise masking can be added to support secure aggregation, but as many clients (numbered more) are involved, O(K^2) pairwise communication channels between them are needed, which is not feasible. Ensuring that all hospitals in a health care network participate in the collection and processing of the data, as well as ensuring compliance with security standards can be challenging, especially for a large-scale network with many hospitals. Other mechanisms, such as

threshold secret sharing & efficient secure aggregation protocols with tree-structured communication topology (Bonawitz et al, 2017) will be explored for enhancements on scalability. In addition and most significantly, computing gradients for each of the training samples in a mini-batch requires a significant amount of computation in deep neural networks with a number of millions parameters. As it is reported in our experiments, small-scale models can be implemented, but if EHR data happens to be longitudinal, a larger model (like a Transformer-based encoder) can be used, which will be a limiting criteria in calculating gradients per-sample.

Thirdly, our framework will be evaluated in practice for its robustness to client dropouts and stragglers – which is a standard setup in any federated deployment in practice. For the experiments, there was no failure for all 10 clients for each communication round. But due to resource constraints at your hospital, or networks going down or needing to be maintained, you may not have the ability to participate. The aggregation model currently implemented is based on the assumption that all the clients participate in all the rounds; then the weighted averaging formula used in fitting the state estimation problem (Equation 13) does not explicitly take into account the presence of clients who are not participating in a round. This is not the entirety and should be complemented with a dropout resistant aggregation mechanism (computing only a weighted average of the clients that did drop out for instance or re-normalizing weights etc), or by employing asynchronous federated learning, meaning that the clients update differently (not necessarily simultaneously). The coupling of the asynchronous updates and the injection of the differential privacy noise is still open research, since the version of the gradient that is considered as a stale version requires a different differential privacy noise parametrization compared to the differential privacy noise parametrization used for simultaneously-available gradients.

Regulatory Alignment and Pathways for Interpretable Federated Intelligence

The laws and regulations around privacy of healthcare data is dynamic; from one law to the next, the requirements on processing this data and training the models become even more stringent, such as in the U.S. with the HIPAA regulations, in Europe with the GDPR, or in China with PIPL. The data minimization principle of the regulations (only model updates are shared – not raw data) and the purpose limitation principle (the global model has been trained only for the intended clinical prediction task) are very relevant to our proposed PP-FedAvg approach. There are a couple the factors that are important to consider though.

As one of the critical issues, the meaning for “personal data” for differentially private model update is unclear. Differential privacy provides answers to the important question of how closely the contribution of any individual patient record must be controlled by a parameter of ε; regulatory agencies may also desire additional assurances that the model itself does not contain identifiable information. In federated learning, for instance, the ask to eradicate the personal data of individual patients under GDPR's right to erasure, e.g. Article 17 (3), implies erasure of local contributions in the global model after training, but this does not seem to readily translate to such a context. This has motivated the research on Federated Machine Unlearning, which seeks to efficiently remove the contribution of some data subset from a learned model, in a manner not too heavy on the model's training. This would help its compliance with regulations, but could be difficult to achieve for unlearning capabilities such as for our PP-FedAvg framework, where the individual contributions are concealed by the DP noise.

An additional regulatory aspect is whether the federated learning process is auditable and transparent. Healthcare organisations or regulators may require proof of privacy assurances provided while training, which could lead to strong privacy accounting mechanisms.To meet this requirement when proving the privacy of the guarantees even during the training, strong privacy accounting solutions are needed. The arguments in support of using the moments accountant are that the moment-endowment (the privacy

budget ε) over multiple rounds, and that this is based on making assumptions about the distribution of data and the clipping that may not be strictly true in reality. The techniques of verifiable privacy auditing to give cryptographic evidence on the amount of privacy budget used may be developed in future work such as using zero-knowledge proofs or trusted execution environments. Moreover, the human understandability of the final model is an important element that can lead towards patient uptake and clinical action adoption: if a model human being has to understand why the model is able to predict this, he/she will be more likely to trust the model/can take action. In addition, federated learning methods with differential privacy introduce another degree of obscurity: the noise introduced together with the randomizations can affect other visibility of how the model predictions are related to the input features.

Conclusion

The Federated learning approach via the Differential privacy technique was proposed in this paper, which has been optimized for the case of high accurate precise prediction in the distributed healthcare institutions without sacrificing privacy. Native federated averaging (FedAvg) is presented here to address this fundamental conflict by concurrently training the models locally, adding Gaussian noise to the model at the end of a certain number of steps, and averaging with the weights. We use a clinical record-at-a-glance approach to keep the raw clinical data within the context of its healthcare institution, and a differential privacy mechanism with a tightly mathematical guarantee that it prevents any inference attack on to any model update shared between institutions.

The experimental results on the two real world datasets (MIMIC-III: predicting hospital outcome for mortality, Diabetes Health Indicators Dataset: predicting diagnosis) demonstrate that the proposed PP-FedAvg framework can predict the mortality event by 3.6-4.2 percentage points ahead of the locally trained models and 0.6-0.7 percentage points ahead of the centralized model. By setting the budget of privacy $\varepsilon = 8$ the framework would keep AUC above 0.93 and only be close after 120 communication rounds. Also the added improvements of 1.1 percentage points with dynamic clipping over fixed clipping in terms of variance removal, and the elimination of convergence time of 35 rounds when using fixed clipping, make dynamic clipping effective in eliminating averaging variance induced by noise.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H., et al. (2016). Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*.

Abbas, S., Abbas, Z., Zahir, A., & Lee, S. (2024). Federated learning in smart healthcare: A comprehensive review on privacy, security, and predictive analytics with IoT integration. *Healthcare*.

Akter, M., Moustafa, N., Lynar, T., et al. (2022). Edge intelligence: Federated learning-based privacy protection framework for smart healthcare systems. *IEEE Journal of Biomedical and Health Informatics*.

Alpan, K., Arman, B., & Dimililer, K. (2024). Performance evaluation and comparison of machine learning algorithms in classification of cdc diabetes health indicators dataset by weka. *2024 8th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*.

Beutel, D., Topal, T., Mathur, A., Qiu, X., et al. (2020). *Flower: A friendly federated learning research framework*. arXiv preprint arXiv:2007.14390.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., et al. (2017). Practical secure aggregation for privacy-preserving machine learning. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*.

Butt, M., Tariq, N., Ashraf, M., Alsagri, H., Moqurrab, S., et al. (2023). A fog-based privacy-preserving federated learning system for smart healthcare applications. *Electronics*.

Chawla, N., Bowyer, K., Hall, L., et al. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*.

Cheon, J., Kim, A., Kim, M., & Song, Y. (2017). Homomorphic encryption for arithmetic of approximate numbers. *Lecture Notes in Computer Science*.

Johnson, A., Pollard, T., Shen, L., Lehman, L., Feng, M., et al. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*.

McMahan, B., Moore, E., Ramage, D., et al. (2017). Communication-efficient learning of deep networks from decentralized data. *Artificial Intelligence and Statistics*.

Murala, D., Krishna, G., Kiran, T., & Mohamud, A. (2025). MedShieldFL-a privacy-preserving hybrid federated learning framework for intelligent healthcare systems. *Scientific Reports*.

Wani, R., & Can, O. (2025). FED-EHR: A privacy-preserving federated learning framework for decentralized healthcare analytics. *Electronics*.

Yousefpour, A., Shilov, I., Sablayrolles, A., et al. (2021). *Opacus: User-friendly differential privacy library in PyTorch*. arXiv preprint arXiv:2109.12298.

Zhao, H., Sui, D., Wang, Y., Ma, L., & Wang, L. (2025). Privacy-preserving federated learning framework for multi-source electronic health records prognosis prediction. *Sensors*.

Ziller, A., Trask, A., Lopardo, A., Szymkow, B., et al. (2021). Pysyft: A library for easy federated learning. *Studies in Computational Intelligence*.