# Privacy-Preserving Predictive Analyticsin Healthcare Using Federated Learningand Deep Learning Models

MScResearchProject
MSc in Data analytics

SricharanPatlori
StudentID:x23231769

SchoolofComputing
NationalCollegeofIreland

Supervisor: Prof.Jorge Basilio

**National College of Ireland**
**Msc in data analytics**
**School of Computing**

| | |
|---|---|
| **Student Name:** | Sricharan Patlori |
| **Student ID:** | x23231769 |
| **Programme:** | Msc in Data analytics |
| **Year:** | 2024-2025 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Jorge Basilio |
| **Submission Due Date:** | 29/1/2025 |
| **Project Title:** | Privacy-Preserving Predictive Analytics in Healthcare Using Federated Learning and Deep Learning Models |
| **Word Count:** | 7150 |
| **Page Count:** | 22 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | Sricharan patlori |
| **Date:** | 29th January 2025 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | □ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Privacy-Preserving Predictive Analytics in Healthcare Using Federated Learning and Deep Learning Models

Sricharan Patlori
x23231769
x23231769@student.ncirl.
ie National College of
Ireland

### Abstract

Federated Learning (FL) is one of the most used strategies to solve the issue of preserving privacy in deep learning applications in healthcare. Centralized models of machine learning are a challenge for data sharing because of the patient privacy and data protection laws which do not allow data sharing between institutions. FL offers a distributed learning solution in which the data is stored in different institutions while sharing the update of the model to ensure that privacy is not infringed while at the same time learning. In this study, a decentralized FL model performed slightly better than standard benchmark models in predicting patient length of stay. Even though centralized models exhibited similar performance like decentralised deep learning model, the objective of this research was not to achieve the highest accuracy possible but to prove that competitive performance can be achieved at the same time as privacy preservation. In the context of healthcare applications, Federated Learning's capability to train models without exchanging the data and yet obtain high accuracy. The motivation for this research is to establish an approach for achieving high accuracy in decentralized systems that also respects privacy and security. The results stress that Federated Learning could be an effective solution that provides both high predictive accuracy and solid protection of privacy-sensitive health data.

## 1 Introduction

The healthcare industry is undergoing a digital transformation, driven by the integration of advanced technologies into patient care, medical research, and operational management. Central to this transformation is the rise of predictive analytics, which uses data to envision the future to determine a patient's prognosis, allocate available resources, and drive clinical choices. The potential of Predictive analysis lies in its ability to handle masses of data collected in healthcare organizations every day, including EHRs, diagnostic images, outputs from wearable devices, and genomic data (Chamikara et al.; 2021; Feng and Buyya; 2016).

However, the use of such data introduces a critical challenge: preserving the confidentiality and the protection of patient's data. Vulnerabilities and threats penetrating the healthcare data are not only infringing on patients' privacy but also causing a myriad of losses in monetary terms and tainted image of the institutions

(Sheller et al.; 2020; Sharma and Guleria; 2023). Meeting both of these objectives, that is developing robust and efficient predictive analytics while protecting data privacy, has become one of the main priorities in the field of healthcare innovation.

*Federated Learning (FL)* is one of the most promising solutions to this challenge. FL is one of the transformative frameworks in machine learning since it enables the training of models across various organizations with none of the data being vulnerable to leaving its local ecosystem. Unlike traditional data aggregation, FL stores the raw data far away from the common learning center (Ziller et al.; 2021; Yang et al.; 2019). However, in a distributed system, only the model updates, like gradient or weights are passed onto the central server. Consequently, it reduces privacy risks while allowing organizations to gain from the collective learning model.

## 1.1 Machine Learning and Deep Learning in Healthcare

Machine learning (ML) and Deep learning (DL) have become innovative tools in the healthcare sector and the ways data is applied and analyzed. *Machine learning*, in a basic understanding refers to a set of algorithms that can learn from data and adapt to enhance performance. The methods include logistic regression, decision trees, and support vector machines which are applied in healthcare settings including risk assessment of patients, diagnosis of diseases, and organizational planning (Li et al.; 2020).

The subset of ML is called *deep learning* where neural networks with a hierarchy of layers are used to analyze datasets for hidden features. DL has brought innovation to the medical imaging area where conventional neural networks have been employed for functions such as tumor detection and organ segmentation. RNNs and their derivatives have been used in time series analysis to improve the future deterioration of patient conditions based on historical health data (McMahan et al.; 2017; Sharma and Guleria; 2023).

These methods have progressed healthcare analytics to a great extent, these approaches of data collection have important privacy issues due to their centralized data gathering process. Although the centralized systems help to collect multiple datasets from various sources, they endanger patient's data and obstruct access to them. This limitation emphasizes the need to adopt privacy-preserving methods such as FL that combines the best of ML and DL without trading data privacy.

## 1.2 Motivation for the Research

The motivation for this research stems from the confluence of several critical factors:

- **Increasing Dependence on Predictive Analytics:** Healthcare companies and organizations globally are incorporating predictive analytics to bring better patient care, cost efficiencies, and more effective resource management. However, the effectiveness of these systems depends on the availability of large, varied, and often disparate datasets that are stored across different organizations.

- **Growing Data Privacy Concerns:** Since the availability of health data is increasing, healthcare organizations need to meet the requirements of the GDPR and HIPAA.

High-profile hacks in health care have shown that centralized systems are very prone to attacks.

- **Underutilization of Federated Learning in Healthcare:** Despite the potential of FL in various application areas, the use of FL in healthcare is not explored significantly. The nature of healthcare datasets as well as their sensitive nature makes it difficult to gather them and train models on them and this is where FL comes in.

This research aims to fill these gaps by showing how FL can revolutionize healthcare analytics. Besides, through decentralizing the training process, FL not only increases privacy but also makes the cooperation of institutions that can be uncomfortable to share data because of competitive or regulatory reasons.

## 1.3 Need for the Research

Despite the potential of FL, several challenges remain unaddressed, making this research timely and necessary:

- **Balancing Privacy and Performance:** Federated learning based models should be capable to achieve balanced performance compared to other models while also preserving privacy without sharing data.

- **Scalability and Robustness:** Whether FL makes sense in real-world healthcare applications when data is from multiple sources and distributed in a rather imbalanced manner, remains an open question.

- **Lack of Real-World Implementations:** However, theoretical research is rich while the actual applications of FL in healthcare especially those that incorporate elaborate DL architectures are scarce.

This study focuses on bridging these gaps by showcasing the practical application of FL combined with DL to address a specific healthcare challenge: This paper aims at establishing the relationship between patient demographic characteristics and length of stay and; The study is not only beneficial for the academic development of FL, but also for clinicians working in the medical field.

This research addresses research question: **”How can Federated Learning based decentralized model can be effectively implemented in healthcare systems to preserve data privacy while maintaining predictive performance compared to centralized machine learning models?”**. This research addresses this question by implementing custom federated learning based deep learning model while also comparing with various centralized machine learning models.

## 1.4 Contributions of the Research

This thesis presents a new solution for privacy-preserving predictive analytics in healthcare based on FL with deep learning models. Key contributions include:

1. **Development of a Privacy-Preserving Framework**: An FL framework is developed to compute an aggregate length of stay for patients across various data sources. The framework maintains strict privacy aspects so that only model updates rather than patient information are exchanged across institutions.

2. **Integration of Advanced Deep Learning Models:** To improve the predictive performance, deep learning structures are used under the FL architecture.

3. **Comparative Analysis of Centralized vs. Federated Models:** FL-based models are compared with traditional centralized model-based approaches based on performance indicators. This analysis shows that, at the cost of more accurate predictions, personal information privacy is compromised by other algorithms.

4. **Simulation of Real-World FL Scenarios:** The study also simulates the actual FL environment since dataset datasets is split in to multiple sources typically like of healthcare organizations. This approach shows that FL is a framework that can be expanded and implemented in various environments.

The findings of this research address a critical gap relating to the balance between privacy and improved predictive accuracy in FL applications for healthcare transformation. The findings contribute not only to the enhancement of academic theory but also offer a case for how safe and feasible analytics can be deployed in practical healthcare organizations.

# 2 Literature Review

## 2.1 Federated Learning in Healthcare

FL (Federated learning) has been studied extensively for healthcare data analysis as a privacy-preserving approach to train machine learning models across decentralized data while keeping data private. Several works have shown that FL can enhance group work while maintaining privacy in different learning methods, especially in activities like disease diagnosis and patient prognosis (Xu et al.; 2021; Rieke et al.; 2020; Pati et al.; 2024). However, to date, FL addresses the privacy issue; the problem is that the application of this approach in actual healthcare settings is still less than ideal due to issues such as data heterogeneity and communication load (McMahan et al.; 2017; Passerat-Palmbach et al.; 2021). Nevertheless, the work proves FL increases predictive accuracy and sustains patients' privacy, making FL a promising strategy for privacy-preserving machine learning in healthcare (Sharma and Guleria; 2023; Cremonesi et al.; 2023). These studies call for future work in extending the integration of FL with deep learning models and practical implementation, especially in large-scale healthcare systems (Antunes et al.; 2022; Li et al.; 2020).

## 2.2 Challenges with Centralized Machine Learning Models

The centralized machine learning models are good at pooling data to enhance the models' performance, they are highly vulnerable to privacy threats and have been observed to leak sensible patient details in various health-related data breaches in the past years (Ziller et al.; 2021; Yang et al.; 2019). The first limitation of these models is the collection of big data into a centralized repository, which increases the vulnerability of the data to leaks and can become a problem when the volume of healthcare data grows (Kairouz et al.; 2021; Rahman et al.; 2023). Centralized approaches also fail to address data heterogeneity and quality issues arising from multiple institutions which if not well balanced, may lead to biased or ineffective models (Hohman et al.; 2020; Sheller et al.; 2020). FL solves these problems by decentralizing both the data and the training process, avoiding these risks, as well as allowing for collaborative model creation while maintaining privacy (Bonawitz et al.; 2019; Chamikara et al.; 2021). However, FL still has a transition from theory to practice as a challenge, and the incorporation of privacy-preserving mechanisms such as differential privacy and secure aggregation is still an active area of research (Ziller et al.; 2021; McMahan et al.; 2017).

## 2.3 Integration of Deep Learning with Federated Learning

Deep learning (DL) when combined with federated learning improves the accuracy of the models, especially in complicated healthcare applications including image analysis and time series prediction (Sharma and Guleria; 2023; Rahman et al.; 2023). The feature of DL to analyze high volumes of data and fine patterns could support FL's privacy consciousness to enable learning without data aggregation (Rieke et al.; 2020; Li et al.; 2020). Nevertheless, this integration brings some issues including the enhanced model complexity of DL models and the additional cost for exchanging model updates (Hossain et al.; 2023; Kairouz et al.; 2021). Several works have suggested the following optimizations for federated DL systems: adaptive learning rates, regularization techniques, and federated averaging (Pati et al.; 2024; McMahan et al.; 2017). However, current innovations in implementing DL in FL frameworks in the healthcare field are still in their infancy and need further investigation for the considerations of model convergence, data distribution, and client selection (Cremonesi et al.; 2023; Antunes et al.; 2022).

Nevertheless, the challenges, including data heterogeneity, communication overhead, and model convergence, which were addressed in federated learning (FL) for healthcare still remain (Xu et al.; 2021; McMahan et al.; 2017). Although FL has privacy benefits over centralized models, the combination of DL with FL is still in its infancy, especially in the more complex healthcare environment (Sharma and Guleria; 2023). Current solutions are based on the theoretical approach and fundamental models and lack integration into practice (Ziller et al.; 2021). This underscores the importance of studying the combination of FL with DL to enhance privacy and prediction at the same time in healthcare because of the scalability and data heterogeneity.

# 3 Data Preprocessing

The dataset applied in this study included 24 variables and 50,000 records with patient data, disease, and hospital characteristics. Preprocessing was crucial for making the dataset optimized for predictive modeling, and for improving both the dataset and the model.

## 3.1 Initial Data Inspection and Cleaning

The dataset was used was checked for data quality like missing values. The review also revealed that there were no missing values for all the corresponding columns. This clean dataset is helpful for further processing thus minimizing the chances of developing models with biases.

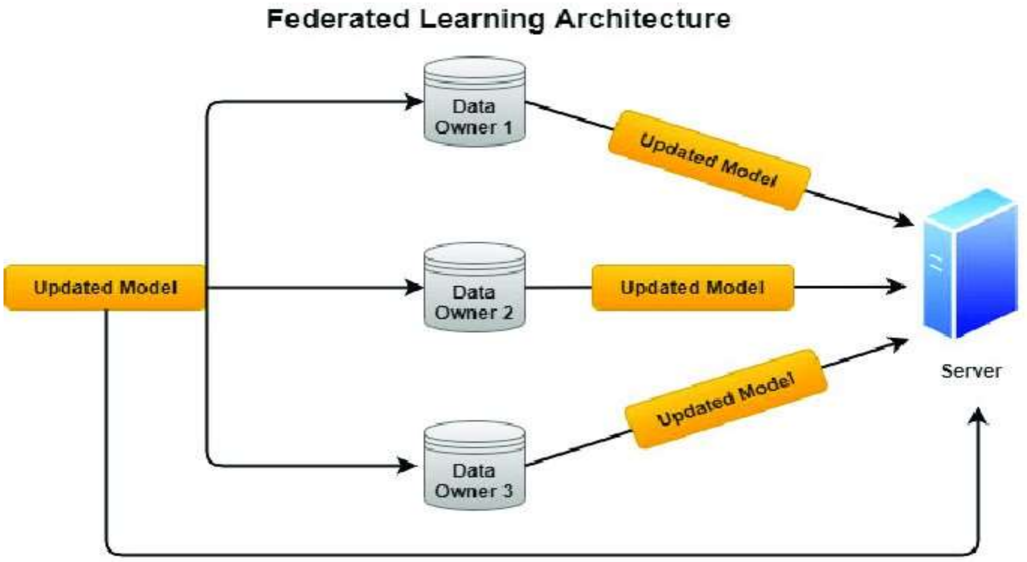


Figure 1: Basic Federated Learning architecture

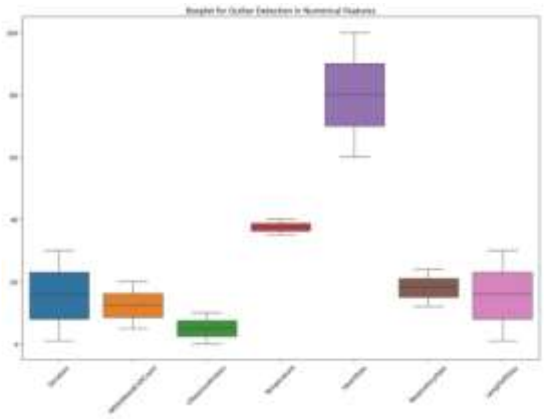


Figure 2: Boxplot for Outlier Detection in Numerical Features

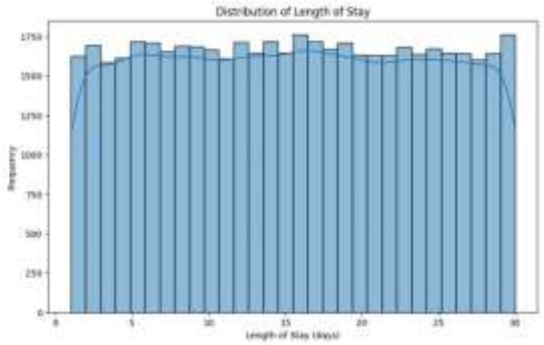


Figure 3: Distribution of Length of Stay

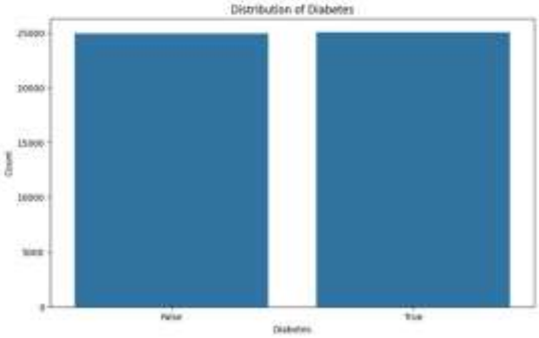


Figure 4: Distribution of diabetes disease

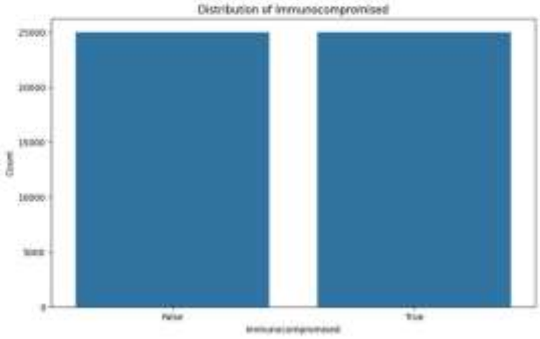


Figure 5: distribution of immunocompromised disease

Figure 6: Various visualisations on dataset

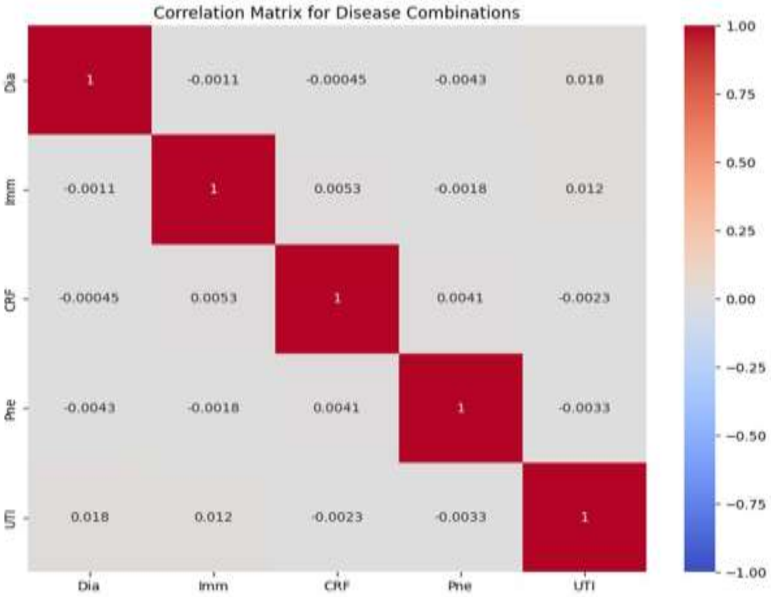


Figure 7: Correlation Matrix for Disease Combinations

## 3.2 Dropping Irrelevant Features

Some of the features are not relevant to the predictive task at hand. The PatientName column which was included only for the identification of patients was dropped as they do not have any correlation with the target variable. The Ethnicity column also was eliminated from the dataset as it does not have much information. The removal of these columns made the dataset simpler to process since the model only had to consider the most important predictors in terms of computational time.

## 3.3 Categorical Data Encoding

Several categorical features needed to be converted to numerical for use in the machine learning models. For example, Age Group was discretized into numerical labels; for the Adult, Elder, and the Infant the values are 0, 1, and 2 respectively. The same applies to other variables such as Gender, CatheterType, and InsertionSite that were also transformed. This process ensures that all categorical data can be read directly by the machine without the loss of interpretability to be easily integrated into the model.

## 3.4 Correlation Analysis

Correlation analysis was conducted to compare the information of different diseases. The study showed that several diseases were inversely related, meaning that the existence of one disease reduced the probability of the other disease. These insights were helpful in feature selection as the model looked at potential interactions between medical variables and narrowed down to the most important predictors while minimizing feature duplication.

## 3.5 Transforming prediction variable for Generalization

First, LengthOfStay was used as a continuous variable, which measures the exact number of days a patient spent in the hospital. To enhance the generalization of this variable for the model, it was sometimes converted into three-day categories by dividing the number of days by three and adding one. For instance, a patient who was admitted and stayed for one to three days was rated as 1, a patient who stayed for four to six days was rated as 2, and so on. Also, the binning strategy enhances the formation of more general models are required in real-time operations. In reality, healthcare settings face many unpredictable and diverse patient populations. Because the model can break down the hospital stays into reasonable time frames, it can offer good estimates in different cases. This makes it suitable for real-time predictions since the results are fast and accurate to help the healthcare system allocate resources, admit patients, and run the healthcare system efficiently.

## 3.6 Outlier Detection

An outlier detection test was performed using the Z-score test to test for any point that is significantly different from the mean. The descriptive statistics showed that there were no extreme observations within the dataset which supported the idea that the distribution of the data was reasonable. This step made the data consistent and minimized cases of having to predict based on outlying values which improves the reliability of the predictive model.

## 3.7 Final Prepared Dataset

After such data preprocessing the dataset was transformed into a clean format suitable for use in machine learning algorithms. The data completeness was addressed, while the features with redundant information were eliminated, categorical data was encoded, correlations were calculated, the target variable was transformed and outliers were detected to provide the dataset amiable for the PP-PA in the healthcare system. Such an approach provided a basis for establishing accurate and reliable machine-learning models.

# 4 Research Methodology

The methodology for this study is to analyze a Federated Learning (FL) approach to train a predictive model on patient length of stay in hospitals while preserving patient privacy. This framework allows for the distributed training of models across multiple institutions that hold patient data while keeping that data within separate institutions. This method is designed and implemented based on federated learning, which distributes computations across clients and collects model updates from a central server.

## 4.1 Model Development and Federated Learning Framework

Federated learning (FL) enables several healthcare facilities (referred to as **clients**) to jointly train a common deep learning model while maintaining data privacy. However, there is no transmission of raw data between clients; each client trains the model only on its dataset and sends updates, such as weights, to a central server. This is decentralized and hence efficient and the data is protected especially in health-related information.

### 4.1.1 Local Training at Each Client

In the FL setup, each client *k* possesses its own dataset $D_k$, which consists of feature-label pairs $\{(x_i,y_i)\}$. The model at client *k*, denoted as $\theta_k$, is locally trained by minimizing the local loss function:

$$\mathcal{L}_k(\theta_k) = \frac{1}{|D_k|} \sum_{(x_i,y_i)\in D_k} \mathcal{L}(f(x_i;\theta_k), y_i) \qquad (1)$$

Where:

- $L(f(x_i;\theta_k),y_i)$ is the loss function for a single training example, measuring the difference between the predicted output $f(x_i;\theta_k)$ and the true label $y_i$.

- $|D_k|$ represents the size of the dataset $D_k$ at client *k*, i.e., the number of data samples used by client *k* for training.

- $\theta_k$ represents the model parameters (e.g., weights and biases) used at client *k*.

In this step, the local model updates, $\Delta\theta_k$, are computed as the difference between the model before and after training:

$$\Delta\theta_k = \theta_{k\text{new}} - \theta_{k\text{old}} \qquad (2)$$

### 4.1.2 Federated Aggregation

Once local training is done, each client *k* sends its model update $\Delta\theta_k$ back to the central server. The server then combines these updates by a **weighted mean** for the weight is proportional to the size of the client dataset. This process is mathematically described by

$$\theta_{\text{global}} = \frac{1}{\sum_{k=1}^{K} |D_k|} \sum_{k=1}^{K} |D_k| \Delta\theta_k \tag{3}$$

Where:

- $\theta_{\text{global}}$ denotes the updated global model after aggregation.
- $K$ is the total number of clients participated in the federated learning process.
- $|D_k|$ denotes the size of the dataset at client $k$, which indicates how many samples are available at that client for training.
- $\sum_{k=1}^{K} |D_k|$ is the total number of training samples across all participating clients, which serves as the normalization factor for the aggregation.
- $\Delta\theta_k$ is the update to the model parameters computed at client $k$.

### 4.1.3 Global Model Update

Once the aggregation step is over, the current global model parameters $\theta_{\text{global}}$ are updated on the server side and then broadcasted for local training in the next round. This process goes on until the model starts nearing a solution and their objective is to enhance the robustness of the model for all the clients while ensuring the data privacy is not compromised in any way.

## 4.2 Implementation Strategy

For federated learning, this research develop a novel deep learning model tailored for LoS (length of stay) prediction with different medical attributes. To enhance the model's ability to generalize the patterns observed in the training phase we incorporate several hidden layers each including dropout and batch normalization to minimize overfitting.

The output layer of the model is a regression layer, allowing it to be used for continuous prediction beneficial for healthcare problems. Actually, the model is trained through several iterations, where in each iteration clients perform local training of the model on their datasets and after that, the server performs federated averaging of the received model updates. The aim is to achieve a global model that performs well in estimating patient LoS in participating institutions and should be more accurate than non-federated models.

### 4.2.1 Federated Training Rounds

Training for the federated learning was done in numerous rounds. Each round consisted of the following steps:

- **Local Training:** Every chosen client $k$ underwent training with their local dataset $D_k$ applying the Adam optimizer as the learning rate was set to adapt to the data.

- **Model Update Transmission:** After the training process was performed, each client transmitted its model update $\Delta w_k$ to the server.

- **Global Aggregation:** The server combined the received updates via the federated averaging formula and updated the weights of the global model.

### 4.2.2 Framework Utilized

TensorFlow Federated (TFF) was used to manage the distributed learning process while leveraging TensorFlow for compatibility and deployment.

- **Training Rounds and Client Participation:** The training process was done in cycles where each cycle consisted of the following steps:
    1. Training on clients' private data with epochs as the number of iterations on the data.
    2. Communication of the locally trained model updates with the server.
    3. Server-side update aggregation by using the federated averaging algorithm to generate a global model.

**Client-Side Optimization:** Clients were initialized with the Adam optimizer which is known for its ability to deal with sparse gradients and its ability to adapt during local training. The local objective was to minimize Mean Squared Error (MSE), expressed as:

$$\mathcal{L}_{\text{local}}(w_k) = \frac{1}{|D_k|} \sum_{i=1}^{|D_k|} (y_i - f(x_i; w_k))^2 \tag{4}$$

where $f(x_i;w_k)$ is the predicted output for input $x_i$ using the model weights $w_k$.

**Server-Side Optimization:** The server used a Stochastic Gradient Descent (SGD) optimizer for summing the client model updates and for updating the global model. This choice helps in the efficient computation in large-scale federated learning systems.

## 4.3 Federated Data Simulation and Preprocessing Strategy

To emulate a realistic federated learning setup:

- **Data Partitioning:** The data set was split into subsets of independent data, each representing data from different clients that mimic the data possessed by different healthcare facilities. The distribution characteristics of the data were maintained in each of the client datasets to offer realistic variability during training.

- **Data Processing at Clients:** Every client was preparing data batches and local training in parallel with the other clients. Local epochs were repeated for batches to mimic several passes through the entire training set for good local model updates.

This methodology focuses on the safe training of predictive models through the federated learning approach, which allows the patient's data to remain stored at the local site while updating the model in a shared manner. Deep learning and federated learning can be integrated to predict the patient length of stay accurately, without violating the privacy laws in healthcare.

# 5 Comparison of Centralized Models

Besides the FL approach, several other traditional machine learning models were also tested and trained to set a benchmark of the predictive performance for the LoS prediction model. Nevertheless, these models are centralized and, although they were used to compare the efficiency of FL, which is privacy-preserving, their application represents an aggregation of the data into a single repository for training. Both models were chosen because these models are effective for regression tasks and for healthcare applications, in particular.

## 5.1 Random Forest

Random Forest is a form of ensemble technique that combines the result of numerous decision trees (Ali et al.; 2012). Every tree is trained with the sample selected randomly from the total data and the final outcome is the sum of the output of all individual trees. This model was selected due to its stability especially when dealing with large features of data as used in this research study on healthcare.

When applied to predicting patient length of stay tasks, Random Forest's best features are its ability to handle non-linearities and interactions between the features, without needing much parameter tuning. The benefit of Random Forest is that it minimizes the problem of overfitting; this is a significant weakness in many feature models. Due to its high scalability and the capacity to include a large number of interactions in the models, it becomes highly useful for healthcare data, as patient states and outcomes depend on numerous factors. Nevertheless, Random Forest means all the data should be collected and processed at the central point which sometimes could lead to critical issues regarding data protection and privacy while implementing in actual healthcare situations.

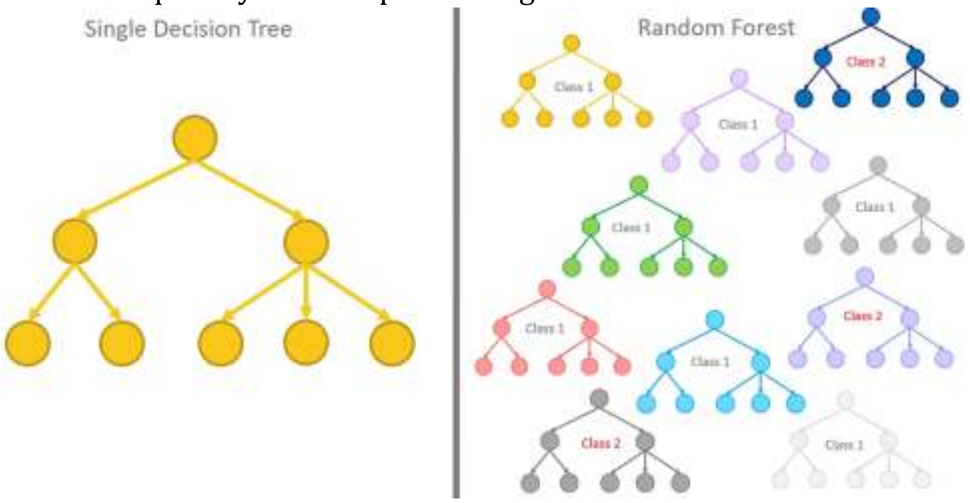


Figure 8: Architecture of Decision Tree and Random Forest(Ali et al.; 2012)

## 5.2 Decision Tree

Decision Trees work in a way that they split the data repeatedly based on feature value, and the chain thus formed is easy to visualize. Both splits try to reduce the prediction error to make the nodes based on the data (De Ville; 2013).tructure that can be easily visualized. Each split attempts to minimize the prediction error by creating nodes that best separate the data (De Ville; 2013).

In the context of forecasting the patient's length of stay, Decision Trees were trained to set a benchmark model. Their interpretability makes them a good fit for healthcare where it is useful to know why a certain prediction was made. Unfortunately, Decision Trees are very sensitive to overfitting, which can become a significant problem in big data sets such as the one used in this research work. They have optimal performance on small data sets but are less effective on large and complex data sets and therefore not ideal for high-dimensional healthcare data sets. Decision Trees also use centralized data which are not very appropriate for use in privacy-preserving methods similar to Random Forest.

## 5.3 Gradient Boosting

Gradient Boosting (Natekin and Knoll; 2013) is an ensemble method where instead of building a single model many models (often decision trees) are built successively and each new model tries to minimize the error made by the previous one. This process refines the model by concentrating on more difficult-to-classify cases and makes it useful for regression types of problems such as estimating the length of stay of patients.

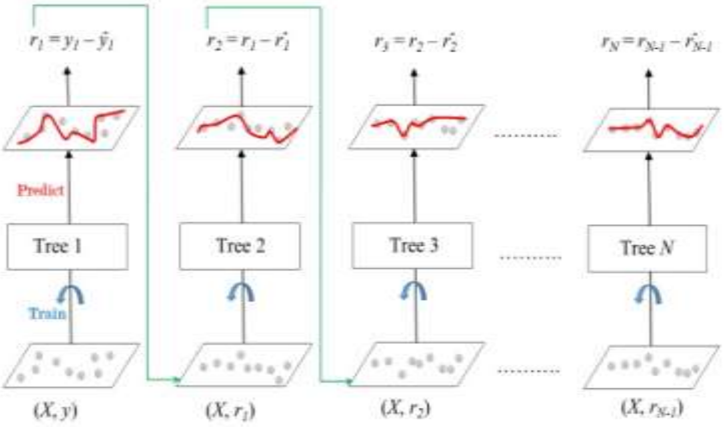


Figure 9: Architecture of Gradient Boosting (Natekin and Knoll; 2013)

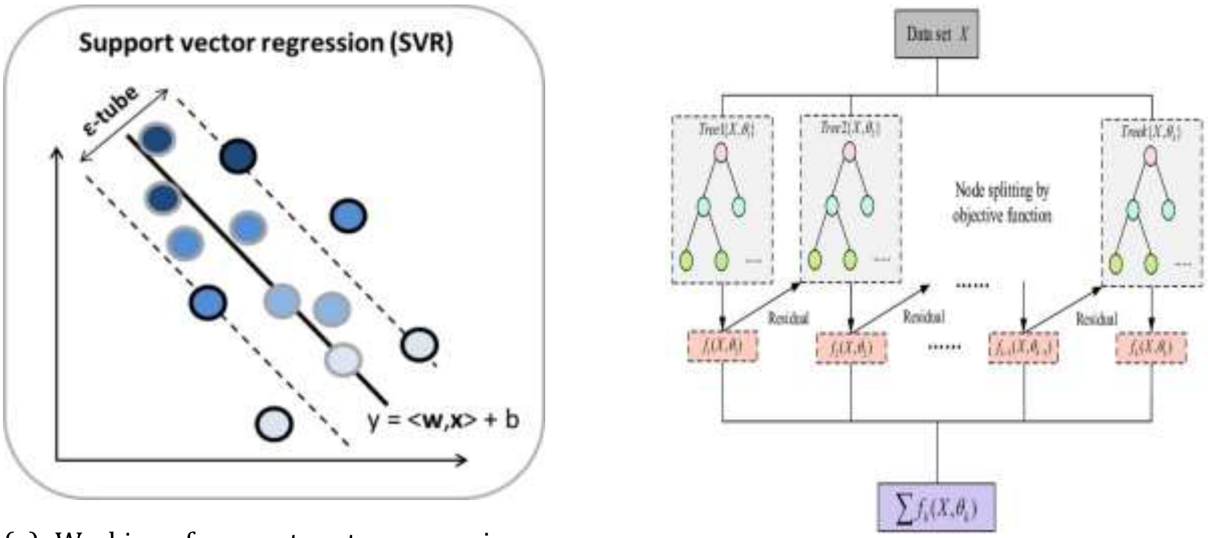
Gradient Boosting also builds the model sequentially to improve the prediction accuracy for the subsequent models, which can be beneficial when predicting healthcare outcomes as often there are multi-factorial dependencies between patient demographics, clinical data, and their medical history. However, the model is complex, and its running time is relatively long, and it may easily overfit if not properly normalized. However, using

Gradient Boosting, this study established that accurate predictions were realized when data was centrally collected. However, it has a drawback in that it depends on the central data store, which raises privacy issues: patient information would have to be transferred between institutions to train the model.

## 5.4 Support Vector Regression (SVR)

SVR (Support Vector Regression) is an extension of the basic concept of SVM which works to identify the best-fit hyperplane. The objective of SVR is to forecast a real value target variable (here- length of stay) up to some error margin. SVR can handle the non-linearity in the features by the use of the kernel functions as pointed out in (Awad et al.; 2015).

The reason for choosing the SVR was because of its ability to handle the high dimensionality of the data and when the features and the target variable interact. It is especially suitable where there are more features than samples within a given set of data. However, SVR can be computationally very expensive, and also the performance of SVR is greatly influenced by the kernel function and kernel parameters to be selected. Furthermore, like the other centralized models, SVR requires the full dataset, which puts the patient's information at risk.



(a) Working of support vector regression (Awad et al.; 2015)

(b) Working of XGBoost (Chen and Guestrin; 2016)

Figure 10: Architecture of various ML models

## 5.5 XGBoost

Extreme Gradient Boosting (XGBoost) is an enhancement of Gradient Boosting to provide even higher speed and performance of the model (Chen and Guestrin; 2016). It also contains methods of reducing overfitting, and therefore it is very efficient to work with big and complicated data sets. XGBoost is one of the most widely used models for structured/tabular data and has been applied in research and practical domains.

XGBoost was chosen because of its performance in predictive tasks as reported by various studies. For these reasons, it is a perfect fit for healthcare predictive modeling because of its strong missing data handling and excellent regularizing capabilities. However, as in all the centralized models, XGBoost brings data into a single repository, which is a major problem when it comes to protecting patients' information in healthcare settings.

### 5.6 Privacy-Preserving Federated Learning

Even though all the centralized models could also show similar performance in terms of predicting patient length of stay, the purpose of this study is not to build the bestperforming predictive model, but a model that functions in a **decentralized** setting while also maintaining patient's privacy.

Federated Learning was presented as a way to overcome such centralized approaches to machine learning. The primary distinction with Federated Learning is that the model is trained **independently** at each participating institution on the local data, and only sends the **model updates** (including gradients or weights) to a central server. This makes the data to be safe and confidential in each university since the raw data is never shared or combined in one location.

The main emphasis of the works is not on achieving the higher precision of the model but on its ability to function in a **federated environment** while preserving the confidentiality of the data. Although in the centralized models, slightly better results is achieved due to the usage of the full dataset at a time, the idea of training a model in a decentralized manner, without sharing data – makes Federated Learning a solution for applications in healthcare that require high privacy. The experiments that are to follow in this study are designed to show that Federated Learning can achieve near similar levels of predictive accuracy as the centralized approach without compromising the privacy of patient data.

## 6 Results and Discussion

| **Model** | **MAE** | **MSE** | **RMSE** |
|---|---|---|---|
| Decision Tree Regressor | 3.3710 | 17.2468 | 4.1529 |
| Random Forest Regressor | 2.5207 | 8.6008 | 2.9327 |
| Gradient Boosting Regressor | 2.8417 | 11.8215 | 3.4382 |
| Support Vector Regression | 2.5288 | 8.7140 | 2.9519 |
| XGBoost Regressor | 2.5500 | 8.9521 | 2.9920 |
| Deep Learning-based Federated Learning | 2.4804 | 8.1507 | 2.8550 |

Table 1: Comparison of various models on various metrics

### 6.1 Evaluation Metrics

Three basic measures including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were employed for the evaluation of the model.

## 6.2 Discussion on ML models

The results of the models showed that the Random Forest Regressor provided the lowest values of MAE, MSE, and RMSE among all the centralized models. The Random Forest model, regarding accuracy, has the following errors: MAE value of 2.5207, MSE value of 8.6008, RMSE value of 2.9327 which is optimal because it balances between accuracy and the time it takes to train. Random Forest because of the way it constructs many decision trees and then combines them was effective in capturing the non-linearity in the healthcare data.

For the Decision Tree Regressor, the accuracy of the model was lower than the Random Forest with an MAE of 3.3710, MSE of 17.2468, and RMSE of 4.1529. While using Decision Trees is quite easy and understanding the results is quite easy, there is a problem with this method: interference, especially when the sample size is large and contains variation. This is expected from Decision Trees, in general, which is a highly overfitting model that leads to high errors for unseen data.

The Gradient Boosting Regresser also gave reasonable results which are an MAE value of 2.8417, an MSE value of 11.8215, and an RMSE value of 3.4382. The Gradient Boosting that creates the new tree model sequentially to rectify the mistakes of the earlier tree model proved the way to learn from complex data. However, it is computationally more expensive and if well tuned has the propensity of overfitting the data and this can explain why it records slightly higher errors than Random Forest.

This proves the SVR model to be reasonable with an MAE value of 2.5288, an MSE value of 8.7140, and an RMSE value of 2.9519. It demonstrates comparatively good results in a condition when it is working in high-dimensional space and the case of nonlinearity of the relation between features and the target variable. Nevertheless, SVR is worse than Random Forest and Gradient boosting since its performance significantly depends on the kernel and hyperparameters, which are more challenging to set compared to ensemble methods.

In addition, the XGBoost Regressor model was also slightly better with an MAE of 2.5500 and MSE of 8.9521 with an RMSE of 2.9920. XGBoost is said to be one of the fastest and most efficient methods of handling large datasets and as such even though the model is comparable to other tree-based models the efficiency that it brings to the largescale application of the model makes it ideal. However, the accuracy results are much worse than Random Forest, and although k-neighboring trees could be chosen, this might be even less accurate due to the use of regularization techniques to avoid overfitting.

## 6.3 Discussion on Deep Learning-based Federated Learning Model

In this study, the proposed Deep Learning-based Federated Learning (FL) model achieved an MAE of 2.4804, MSE of 8.1507, and RMSE of 2.8550, which was marginally better than the traditional centralized models such as Random Forest and Gradient Boosting in terms of predictive accuracy. Even though the results achieved by the model can be compared with centralized models, this research was aimed not only at optimizing the accuracy of the solution but also at creating a solution that works efficiently in a decentralized environment and meets the requirements of data protection. This approach tackles a major problem in the healthcare management domain, that is, patient data is normally shared across different healthcare institutions, but due to considerations of data sensitivity, data sharing is usually limited.

### 6.4 Results of Centralized Models vs. Federated Learning

Compared with other models, Random Forest, Gradient Boosting, and XGBoost generally show better results because they are centralized models and can use the entire dataset. These models advantage of the accumulating of data from many institutions into one data warehouse so that they can have a better and more detailed pattern. However, this advantage is not desirable in the present context of healthcare data because the privacy of patients is one of the most important aspects. Centralization models mean that patient information needs to be transferred from one institution to another, which creates a chance of privacy violation, data misuse, and non compliance with data protection rules like HIPAA or GDPR.

### 6.5 Federated Learning's Key Advantage: Privacy Preservation

The primary advantage of Federated Learning is that training can happen without sharing data with others. As for Federated Learning, all institutions learn their model independently based on their data and send information updates, e.g., gradients or weights, to the central server. This means that the ability to protect patient information is maintained and secured all the time as learners. The model can be trained in the decentralized data sources and the raw data never leaves the institution making the model a privacypreserving solution for healthcare analytics.

Although, when compared to the best-centralized models, Federated Learning may not always produce the highest accuracy, its privacy-preserving nature makes it indispensable for practical healthcare applications. The practice of never exchanging patient data between institutions of Federated Learning makes it possible to build accurate, predictive models without ever violating the patient's right to privacy and data protection.

## 7 Conclusion and Future Work

### 7.1 Conclusion

This study showed that Deep Learning-based FL was slightly superior to other traditional centralized models such as Random Forest, Gradient Boosting, and XGBoost in predicting patient length of stay with slightly lower MAE, MSE, and RMSE than the other models. While centralized models are efficient in learning intricate patterns as they work with the whole dataset, the main benefit of Federated Learning is the capability of training with minimal error while keeping the data privacy. Federated Learning is thus a middle ground between high predictive results and privacy preservation because it allows training to occur in a decentralized manner, with data remaining at the source institution. This approach proves particularly useful in healthcare, where data privacy is of utmost importance, as it shows that Federated Learning is a feasible solution for privacy-constrained applications while also delivering high predictive performance.

## 7.2 Future Work

The future directions of Federated Learning can be related to its improvement in terms of performance of various netrics by investigating various approaches to privacy-preserving methods including secure aggregation and differential privacy. There is also a need to enhance the communication between the involved institutions to improve the time taken to train the model. Another key issue to be addressed will be the problem of how to manage heterogeneity of data collected from different institutions so that Federated Learning can learn across the heterogeneity. But adding more data such as the patient's history, temporal data or other data could help in generating better predictions. Finally, increasing privacy protection measures by applying more elaborate encryption techniques could improve the security of Federated Learning, and safeguard sensitive healthcare information from unauthorized access while implementing efficient decentralized model learning.

# References

Ali, J., Khan, R., Ahmad, N. and Maqsood, I. (2012). Random forests and decision trees, *International Journal of Computer Science Issues (IJCSI)* **9**(5): 272.

Antunes, R. S., Stoffel, C. A., Ku¨derle, A., Abdullahi Yari, I. and Eskofier, B. (2022). Federated learning for healthcare: Systematic review and architecture proposal, *ACM Transactions on Intelligent Systems and Technology* **13**(4): 1–23.

Awad, M., Khanna, R., Awad, M. and Khanna, R. (2015). Support vector regression, *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* pp. 67–80.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivkin, N., Kiddon, C., Konecny, J., Mazzocchi, S. and McMahan, B. (2019). Towards federated learning at scale: System design, **1**: 374–388.

Chamikara, M. A. P., Bertok, P., Khalil, I., Liu, D. and Camtepe, S. (2021). Privacypreserving distributed machine learning with federated learning, *Computer Communications* **171**: 112–125.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Cremonesi, F., Planat, V., Kalokyri, V., Kondylakis, H., Sanavia, T., Resinas, V. M. M., Singh, B. and Uribe, S. (2023). The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform, *Journal of Biomedical Informatics* **141**: 104338.

De Ville, B. (2013). Decision trees, *Wiley Interdisciplinary Reviews: Computational Statistics* **5**(6): 448–455.

Feng, G. and Buyya, R. (2016). Maximum revenue-oriented resource allocation in cloud, *IJGUC* **7**(1): 12–21.

Hohman, F., Wongsuphasawat, K., Kery, M. B. and Patel, K. (2020). Understanding and visualizing data iteration in machine learning.

Hossain, M. S., Muhammad, G. and Kundu, D. (2023). Federated learning in smart healthcare systems: Concepts and applications, *IEEE Internet of Things Journal* .

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G. and Cummings, R. (2021). Advances and open problems in federated learning, *Foundations and Trends® in Machine Learning* **14**(1–2): 1–210.

Li, T., Sahu, A. K., Talwalkar, A. and Smith, V. (2020). Federated learning: Challenges, methods, and future directions, *IEEE Signal Processing Magazine* **37**(3): 50–60.

McMahan, B., Moore, E., Ramage, D., Hampson, S. and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data, pp. 1273–1282.

Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial, *Frontiers in neurorobotics* **7**: 21.

Passerat-Palmbach, J., Ziller, A. and Ryffel, T. (2021). Federated learning systems: Towards next-generation ai.

Pati, S., Kumar, S., Varma, A., Edwards, B., Lu, C., Qu, L., Wang, J. J., Lakshminarayanan, A., Wang, S. H., Sheller, M. J. and Chang, K. (2024). Privacy preservation for federated learning in health care, *Patterns* **5**(7): 100059.

Rahman, A., Hossain, M. S., Muhammad, G., Kundu, D., Debnath, T., Rahman, M., Khan, M. S. I., Tiwari, P. and Band, S. S. (2023). Federated learning-based ai approaches in smart healthcare: Concepts, taxonomies, challenges, and open issues, *Cluster Computing* **26**(4): 2271–2311.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K. and Ourselin, S. (2020). The future of digital health with federated learning, *NPJ Digital Medicine* **3**(1): 1–7.

Sharma, S. and Guleria, K. (2023). A comprehensive review on federated learning-based models for healthcare applications, *Artificial Intelligence in Medicine* **146**: 102691.

Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Bakas, S. and Marcus, D. S. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data, *Scientific Reports* **10**(1): 1–12.

Xu, J., Glicksberg, B. S., Su, C., Walker, P., Bian, J. and Wang, F. (2021). Federated learning for healthcare informatics, *Journal of Healthcare Informatics Research* **5**(1): 1– 19.

Yang, Q., Liu, Y., Chen, T. and Tong, Y. (2019). Federated machine learning: Concept and applications, *ACM Transactions on Intelligent Systems and Technology* **10**(2): 1– 19.

Ziller, A., Trask, A., Lopardo, A., Szymkow, B., Wagner, B. and Bluemke, E. (2021). Pysyft: A library for easy federated learning, pp. 111–139.