Federated Heart-Care using Differential Privacy

A PROJECT REPORT

Submitted in partial fulfilment of the

requirement for the award of the degree

of

BACHELOR OF TECHNOLOGY (B.Tech)

in

Information Technology

by

Yash verma

219302215

(Information Technology)

MANIPAL UNIVERSITY JAIPUR

JAIPUR-303007

RAJASTHAN, INDIA

June 2025

Manipal University Jaipur

MANIPAL UNIVERSITY
JAIPUR
(University under Section 2(f) of the UGC Act)

Date:

CERTIFICATE

This is to certify that the project titled **Federated Heart-Care using Differential Privacy** is a record of the bonafide work done by **Yash verma (219302215)** submitted in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology (B.Tech) in **Information Technology of Manipal University Jaipur, during the academic year 2024-25.**

Dr. Nirmal Kumar Gupta
Project Guide, Dept of Information Technology
Manipal University Jaipur

Dr. Pratistha Mathur
HOD, Dept of Information Technology
Manipal University Jaipur

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my Project Guide **Dr. Nirmal Kumar Gupta**, Assistant Professor, Dept. of Information Technology**,** Manipal University Jaipur for their invaluable guidance, continuous support, and encouragement throughout the duration of this project and research. Their insights, constructive feedback, and patience greatly contributed to the success of this work.

I am profoundly thankful to **Dr. Pratistha Mathur,** HOD, Dept. of Information Technology**,** Manipal University Jaipur for their constant encouragement, administrative support, and for providing all the facilities necessary for the successful completion of this study.

I am also thankful to **Manipal University Jaipur** for providing the necessary resources and a conducive environment to carry out my research. My heartfelt thanks go to all the faculty members and staff for their assistance and support.

I extend my special thanks to my family and friends for their unwavering support, understanding, and motivation throughout my academic journey.

ABSTRACT

Heart disease remains one of the leading causes of mortality worldwide, making early and accurate predictions critical for timely intervention and improved patient outcomes. In today's data-driven healthcare environment, leveraging machine learning to analyze patient data can significantly enhance diagnostic accuracy. However, concerns about patient privacy and data security pose major challenges, especially when data is distributed across multiple healthcare institutions. This project, **Federated Heart-Care using Differential Privacy**, aims to develop a robust, privacy-preserving predictive model for heart disease by combining federated learning with differential privacy techniques, ensuring both accuracy and confidentiality in collaborative medical data analysis.

The methodology involves training multiple machine learning models—**Support Vector Machines (SVM), Logistic Regression, Random Forest**, and **Artificial Neural Networks (ANN)—**within a federated learning framework. This approach allows decentralized training on data from various sources without sharing raw data, thereby maintaining data privacy. Differential Privacy is applied to the model updates to add noise and protect sensitive patient information further. **Exploratory Data Analysis (EDA**) and visualization techniques are employed to understand data patterns and improve model performance. The entire system is designed to provide accurate heart disease risk predictions while safeguarding individual privacy.

The results demonstrate that the federated learning models, enhanced with differential privacy, achieve high predictive accuracy comparable to centralized models, while effectively mitigating privacy risks. This balance between performance and privacy is crucial in healthcare applications, where data sensitivity is paramount. The project highlights the feasibility of secure, collaborative AI-driven diagnostics that can be deployed across institutions without compromising patient confidentiality, paving the way for broader adoption of privacy-preserving machine learning in medical research.

The project is implemented using **Python** and several powerful libraries and tools. **Scikit-learn (Sklearn)** supports the development of classical machine learning models, while Flower facilitates federated learning. **Pandas** and **NumPy** handle data manipulation and numerical computations during EDA, with Matplotlib and Seaborn used for insightful data visualization. **Streamlit** is employed to build an interactive and user-friendly web-based interface, enabling healthcare professionals to access and interpret model predictions in real-time. This comprehensive technology stack ensures a scalable, efficient, and privacy-compliant solution for heart disease prediction.

LIST OF TABLES

LIST OF FIGURES

Contents

Chapter 1
Introduction

1.1 Introduction to Project

Project Federated Heart-Care is an innovative initiative aimed at advancing heart disease prediction by leveraging the combined strengths of federated learning and differential privacy. In the modern healthcare landscape, the digitization of patient data has surged, offering unprecedented opportunities for machine learning to enhance diagnosis, treatment, and patient outcomes. However, this growth also raises significant concerns about the privacy and security of sensitive health information. Federated Heart-Care addresses these challenges by enabling multiple healthcare institutions to collaboratively train predictive models on decentralized datasets without sharing raw patient data, thereby preserving data privacy and complying with stringent regulatory standards.

Federated learning, the core technology behind Federated Heart-Care, facilitates decentralized model training where each participating client (e.g., hospital or clinic) trains a local model on its own data and only shares model updates with a central server. This approach eliminates the need for aggregating sensitive data in a single repository, reducing the risk of data breaches and unauthorized access. To further strengthen privacy protection, the project integrates differential privacy, a mathematical framework that introduces carefully calibrated noise to the data or model parameters. This ensures that individual patient information cannot be re-identified from the shared model updates, providing formal privacy guarantees while maintaining high model utility.

The Federated Heart-Care system is implemented using the Flower framework, which supports efficient and scalable federated learning. It employs advanced aggregation algorithms such as Federated Averaging (FedAvg) to combine local model updates into a global predictive model. This collaborative learning process enhances the accuracy of heart disease prediction models by harnessing diverse datasets from multiple sources while respecting patient confidentiality. Experimental results have demonstrated that the integration of federated learning with differential privacy can achieve robust predictive performance, with test accuracy around 96%, without compromising the privacy of individual data contributors.

Overall, Project Federated Heart-Care exemplifies a cutting-edge approach to healthcare analytics that balances the need for data-driven insights with the imperative of privacy preservation. By uniting federated learning and differential privacy, the project paves the way for secure, collaborative medical research and improved clinical decision-making. This paradigm not only mitigates risks associated with data centralization and cyberattacks but also fosters trust among healthcare stakeholders, ultimately contributing to better patient care and outcomes in the fight against heart disease.

1.2 Background and Motivation

Project Federated Heart-Care is motivated by the critical need to leverage the vast and growing volumes of sensitive healthcare data for improving heart disease prediction while rigorously preserving patient privacy. Healthcare data, including electronic health records, medical images, and real-time monitoring data, are typically distributed across multiple institutions and protected by strict regulations such as HIPAA and GDPR. These regulations restrict direct data sharing, making traditional centralized machine learning approaches infeasible due to the risks of data breaches, unauthorized access, and non-compliance. Consequently, there is a pressing demand for privacy-preserving collaborative learning frameworks that enable multi-institutional model training without compromising sensitive data.

Federated learning (FL) addresses this challenge by allowing multiple healthcare providers to collaboratively train a shared predictive model without exchanging raw patient data. Instead, each institution trains a local model on its own data and shares only model updates (e.g., gradients or weights) with a central server, which aggregates them to update a global model. This decentralized approach mitigates many privacy risks associated with data centralization and reduces the regulatory burden. However, FL alone is vulnerable to privacy leakage through model updates, which can potentially be exploited to infer sensitive information about individual patients or institutions. This vulnerability motivates the integration of additional privacy safeguards.

To enhance privacy guarantees, Project Federated Heart-Care incorporates differential privacy (DP) into the federated learning process. Differential privacy introduces controlled statistical noise to the model updates before sharing, providing mathematically quantifiable privacy protection by ensuring that the contribution of any single data point cannot be distinguished. This mechanism prevents adversaries from re-identifying individual patients from the shared parameters, even if they have access to the model updates. The combination of FL and DP thus balances the trade-off between model utility and privacy, enabling robust heart disease prediction models that comply with stringent data protection laws while maintaining high accuracy.

Technically, the project leverages state-of-the-art federated optimization algorithms such as Federated Averaging (FedAvg) and employs noise calibration strategies to optimize the privacy budget in differential privacy. It also addresses challenges related to heterogeneous and non-IID (non-independent and identically distributed) data across institutions, communication efficiency, and secure aggregation to prevent information leakage during model update exchanges. By integrating these advanced privacy-preserving techniques, Project Federated Heart-Care aims to unlock the potential of distributed healthcare data, fostering secure, scalable, and collaborative AI-driven heart disease diagnostics that respect patient confidentiality and regulatory compliance.

1.3 Applications and Advantages

Project Federated Heart-Care offers significant applications and advantages in the realm of cardiac healthcare. It enables accurate heart disease prediction by collaboratively training models on decentralized patient data from multiple healthcare institutions, thus overcoming data silos and regulatory barriers. The system supports personalized care, real-time monitoring through IoT devices, and integration with clinical decision support systems, enhancing both diagnosis and treatment. Its privacy-preserving design, combining federated learning with differential privacy, ensures compliance with strict data protection laws while safeguarding sensitive patient information. Additionally, the project improves model generalizability by leveraging diverse datasets, enhances communication efficiency, and fosters trust among stakeholders through secure data handling. These features collectively make Federated Heart-Care a scalable, secure, and effective solution for advancing cardiovascular health outcomes.

1.3.1 Applications

1.3.1.1 Heart Disease Prediction:

The project enables accurate prediction of heart disease by leveraging decentralized Electronic Health Records (EHRs) combined with IoT-generated health data, improving early diagnosis and patient outcomes without compromising data privacy.

1.3.1.2 Collaborative Healthcare Analytics:

Facilitates multi-institutional collaboration where hospitals and clinics can jointly train predictive models on distributed datasets, overcoming data silos and regulatory restrictions on data sharing.

1.3.1.3 Personalized Patient Care:

Supports personalized treatment and lifestyle recommendations by integrating multimodal data such as cardiac images, ECG signals, and patient records, enhancing clinical decision-making.

1.3.1.4 Clinical Decision Support Systems (CDSS):

Can be integrated with platforms like mPower Health to empower healthcare providers with evidence-based management plans, risk scoring, and continuous patient monitoring in a privacy-preserving manner.

1.3.1.5 Real-time Monitoring and Remote Healthcare:

Enables secure use of wearable and IoT devices for continuous cardiac health monitoring, allowing timely interventions without exposing sensitive data.

1.3.2 Advantages

1.3.2.1 Data Privacy Preservation:

By combining federated learning with differential privacy, the project ensures that no raw patient data leaves local institutions, and model updates are protected against re-identification attacks, complying with HIPAA and GDPR standards.

1.3.2.2 Scalability and Efficiency:

The use of federated averaging (FedAvg) and optimized algorithms allows efficient training across multiple decentralized nodes, handling high-dimensional healthcare data with reduced communication overhead.

1.3.2.3 Improved Model Generalizability:

Training on diverse datasets from multiple institutions leads to more robust and generalizable heart disease prediction models that perform well across different populations and clinical settings.

1.3.2.4 Enhanced Security and Trust:

Secure aggregation and noise addition via differential privacy build trust among participating institutions and patients by mitigating risks of data breaches and adversarial inference.

1.3.2.5 Support for Multimodal Data Integration:

The framework can fuse heterogeneous data types (images, signals, records) using attention-based models, improving diagnostic accuracy beyond traditional single-modality approaches.

1.3.2.6 Empowerment of Healthcare Workforce:

Through integration with digital health platforms, non-physician healthcare workers can be enabled to deliver quality cardiac care supported by AI-driven insights, even in resource-limited settings.

These applications and advantages collectively position Project Federated Heart-Care as a transformative approach in privacy-preserving, collaborative cardiac healthcare analytics, driving improved patient outcomes and advancing digital health innovation.

1.4 Project Objectives

1.4.1 Develop a Privacy-Preserving Federated Learning Framework

- Design and implement a federated learning architecture that enables multiple healthcare institutions to collaboratively train heart disease prediction models without sharing raw patient data.
- Ensure seamless integration with existing healthcare IT infrastructures to facilitate easy adoption.

1.4.2 Integrate Differential Privacy Mechanisms

- Incorporate differential privacy techniques to add calibrated noise to model updates, providing strong mathematical guarantees against data leakage.
- Optimize the privacy-utility trade-off to maintain high predictive accuracy while safeguarding patient confidentiality.

1.4.3 Enhance Model Accuracy and Generalizability

- Utilize diverse, multi-institutional datasets to train robust heart disease prediction models that generalize well across different populations and clinical settings.
- Implement advanced aggregation algorithms like Federated Averaging (FedAvg) to effectively combine local model updates.

1.4.4 Support Multimodal Data Fusion

- Develop methods to integrate heterogeneous data types such as ECG signals, medical images, and electronic health records using attention-based deep learning models.
- Improve diagnostic performance by leveraging complementary information from multiple data modalities.

1.4.5 Ensure Scalability and Communication Efficiency

- Optimize communication protocols to reduce bandwidth usage and latency during federated training across geographically distributed nodes.
- Design scalable solutions capable of handling large numbers of participating healthcare institutions and high-dimensional data.

1.4.6 Facilitate Clinical Integration and Usability

- Integrate the federated heart-care models with clinical decision support systems and digital health platforms like mPower Health.
- Provide user-friendly interfaces and actionable insights to empower healthcare professionals and non-physician workers in delivering quality cardiac care.
- Conduct pilot studies and real-world validations to assess clinical impact and usability.

Chapter 2
Background Materials

2.1 Conceptual Overview

Project Federated Heart-Care is designed to revolutionize cardiac healthcare analytics by enabling collaborative, privacy-preserving machine learning across multiple healthcare institutions. At its core, the project combines federated learning (FL) with differential privacy (DP) to build robust heart disease prediction models without requiring the centralization of sensitive patient data. Instead of pooling raw data, each participating hospital or clinic trains a local model on its own dataset. These locally trained models then share only encrypted or noise-perturbed updates with a central server, which aggregates them to form a global model. This approach ensures that patient data remains securely within the institution's premises, addressing privacy concerns and regulatory constraints.

The conceptual framework involves several key components: decentralized data sources, local model training, secure communication protocols, privacy-preserving mechanisms, and global model aggregation. Data sources include diverse modalities such as electronic health records (EHR), ECG signals, medical imaging, and wearable device data, which are often heterogeneous and non-IID (non-independent and identically distributed). To handle this complexity, the project employs advanced deep learning architectures, including attention-based models, to effectively fuse multimodal data and extract meaningful features for heart disease prediction.

Differential privacy is integrated into the federated learning pipeline by adding calibrated noise to the model updates before they are shared, providing formal mathematical guarantees that individual patient information cannot be reverse engineered from the aggregated model. This ensures compliance with healthcare data protection regulations like HIPAA and GDPR. Additionally, secure aggregation protocols prevent any single party, including the central server, from accessing raw updates, further enhancing privacy and security.

Overall, Project Federated Heart-Care conceptualizes a scalable, efficient, and secure ecosystem for collaborative cardiac healthcare analytics. By balancing the trade-offs between data utility and privacy, it aims to improve predictive accuracy, foster trust among healthcare providers, and ultimately contribute to better patient outcomes through early diagnosis, personalized treatment, and continuous monitoring, all while maintaining stringent privacy standards.

2.2 Technologies used

Project Federated Heart-Care utilizes a comprehensive suite of modern technologies to deliver a robust, privacy-preserving, and interactive heart disease prediction platform. The core application is developed in Python 3.11, leveraging its extensive ecosystem for data science and machine learning. The user interface is built with Streamlit, allowing real-time data upload, visualization, and model interaction in a web environment. For data preprocessing and analysis, libraries such as Pandas and NumPy are used for data manipulation, while Matplotlib and Seaborn provide advanced visualization capabilities, including distribution plots and correlation heatmaps.

On the modeling front, the project employs classical machine learning algorithms—Logistic Regression, Support Vector Machine (SVM), and Random Forest—from scikit-learn for baseline predictive analysis, as well as SMOTE from imbalanced-learn to address class imbalance in the dataset. For deep learning, PyTorch is used to construct and train Artificial Neural Networks (ANNs), both in centralized and federated settings. The federated learning simulation is implemented using custom logic, with FedAvg (Federated Averaging) for model aggregation across simulated clients (hospitals). Privacy is further enhanced through Opacus, a library that brings differential privacy to PyTorch models by adding calibrated noise to gradients during training.

The project's workflow includes data loading and cleaning, exploratory data analysis, dataset balancing, model training and evaluation, and federated learning simulation with differential privacy. Additional technologies like Torch DataLoader and TensorDataset enable efficient batch processing during model training. This integrated technology stack ensures that Project Federated Heart-Care delivers a scalable, secure, and user-friendly solution for collaborative cardiac healthcare analytics, supporting both rigorous data privacy and high predictive performance.

The key management technologies that have been used in the project Federated Heart-Care Using Differential Privacy are as follows: -

- **Python 3.11**: The primary programming language for the entire project.
- **Streamlit**: Used for building the interactive web-based user interface, allowing for real-time data upload, visualization, and model interaction.
- **Pandas and NumPy:** For data manipulation, preprocessing, and numerical operations1.
- **Matplotlib and Seaborn:** Employed for data visualization, including plotting distributions, heatmaps, and feature importance.
- **Scikit-learn (sklearn):** Utilized for classic machine learning models (Logistic Regression, SVM, Random Forest), data preprocessing (StandardScaler), model evaluation (accuracy, confusion matrix), and train-test splitting.
- **Imbalanced-learn (imblearn):** Specifically, the SMOTE technique is used for balancing the dataset to address class imbalance.
- **PyTorch:** The deep learning framework used for building, training, and evaluating Artificial Neural Networks (ANNs) both in centralized and federated settings.
- **Opacus:** A library that brings differential privacy to PyTorch models, enabling privacy-preserving training by adding calibrated noise to gradients during model updates.

- **Federated Learning Algorithms:** Custom implementation of federated learning simulation, including FedAvg (Federated Averaging) for aggregating model parameters across simulated clients/hospitals.
- **Torch Data Utilities:** Including DataLoader and TensorDataset for efficient data batching and handling during model training.

2.3 Prerequisites

2.3.1 Programming and Analytical Skills

- **Python Proficiency:**

  Strong command of Python (recommended version 3.11) for implementing data processing, machine learning, deep learning, and federated learning workflows.

- **Data Analysis:**

  Ability to handle, preprocess, and analyze data using libraries such as Pandas and NumPy, including encoding categorical variables, handling missing values, and performing exploratory data analysis (EDA) with Matplotlib and Seaborn.

- **Machine Learning:**

  Understanding of classical supervised learning algorithms (Logistic Regression, SVM, Random Forest) and their implementation using scikit-learn, including model evaluation with metrics like accuracy and confusion matrix.

- **Deep Learning:**

  Experience in building and training neural networks using PyTorch, including defining custom architectures, training loops, and evaluation on test data.

- **Imbalanced Data Handling:**

  Knowledge of techniques such as SMOTE (from imbalanced-learn) for balancing class distributions in medical datasets.

- **Federated Learning Concepts:**

  Familiarity with federated learning principles, including local training, model aggregation (FedAvg), and simulation of multiple clients/hospitals.

- **Differential Privacy:**

  Understanding of privacy-preserving machine learning, specifically the integration of differential privacy using the Opacus library in PyTorch to add noise and limit gradient norms during training

2.3.2 Software and Libraries

- **Python 3.11** (or compatible version)

- **Streamlit:** For building interactive web-based user interfaces and visualizations.

- **Pandas, NumPy:** For data manipulation and preprocessing.

- **Matplotlib, Seaborn:** For data visualization and EDA.

- **Scikit-learn**: For classical ML models, preprocessing (StandardScaler), and metrics.
- **Imbalanced-learn (SMOTE):** For oversampling and balancing datasets.
- **PyTorch:** For deep learning model development, training, and evaluation.
- **Opacus:** For implementing differential privacy in PyTorch models.
- **Torch Data Utilities**: Including DataLoader and TensorDataset for efficient data batching and management.

2.3.3 System and Hardware Requirements

- **Operating System:**
  Windows, macOS, or Linux.
- **Hardware:**
  At least 8GB RAM (16GB recommended for deep learning tasks). A CUDA-capable GPU is beneficial for accelerating PyTorch training but not mandatory for small or medium datasets.

2.3.4 Data Requirements

- **Dataset Format:**
  The application supports heart disease risk datasets in both CSV and Excel formats (.csv, .xlsx). Users can upload their own files or use the default dataset provided.
- **Data Preparation:**
  Ensure categorical variables are encoded as integers (e.g., Yes/No to 1/0), and missing values are addressed before model training. The application includes preprocessing steps for these tasks

Chapter 3
Methodology

3.1 Working Methodology

The aim of the working methodology in this project is to develop a robust, privacy-preserving heart disease risk prediction system that leverages federated learning and differential privacy. By enabling multiple healthcare institutions (simulated as clients) to collaboratively train machine learning models without sharing raw patient data, the methodology ensures both high predictive performance and strong data confidentiality. The approach integrates advanced data science techniques, including data balancing, feature engineering, and neural network modeling—with rigorous privacy mechanisms, allowing for secure aggregation of model updates and compliance with data protection standards before any data preprocessing or transformation begins

3.1.1 Data Preprocessing and Transformation

- Dataset: $\{(x_i, y_i)\}_{i=1}^{N}$ **, where** $x_i \in R^d$ **,** $y_i \in \{0,1\}$**.**
- Categorical to Binary Mapping: For categorical feature c, define mapping f_c : category → {0,1}, e.g., f_c("yes") = 1, f_c("no") = 0.
- Missing Value Removal: $D' = D \setminus \{(x_i, y_i):]_j, x_{ij} = NaN\}$.

3.1.2 Data Balancing:

- **Smote Oversampling:** For minority classs y = 1, generate synthetic samples $x_{new} = x_i + \lambda(x_j - x_i)$, $\lambda \sim U(0,1)$, $x_i, x_j \in$ **minority class.**

3.1.3 Centralized Baseline Model Construction

- **Feature Scaling**
  - **Standardization:** For each feature k, $x'_{ik} = \frac{(xik - \mu k)}{\sigma k}$, where μ_k and σ_k mean and std-dev of feature k.

3.1.4 Model Choices

- **Logistic Regression:** $y' = \sigma(w^T x + b)$, $\sigma(z) = 1/1+e^{-z.}$
- **SVM:** Find w, b minimizing $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \max(0, 1 - yi(W^Txi + b))$.
- **Random Forest:** Ensemble of T decision trees, output by majority vote.

3.1.5 Artificial Neural Network (ANN)

- **Architecture:** Input dd, Hidden layers: 32, 16, Output: 1 (sigmoid).
- **Forward Pass:**

  $$h1 = ReLU(W_1 x + b_1),\ h2 = ReLU(W_2 h_1 + b_2),\ y' = \sigma(W_3 h_2 + b3)$$

- **Loss**: Binary cross-entropy

  $$L(y, y') = -[y \log y' + (1-y)\log(1-y')]$$

- **Optimization:** Adam optimizer, learning rate $\alpha = 0.001$.

3.1.6 Federated Learning System Design

- **Client Data Partitioning**
  - **Partitioning: D** = $\cup_{k=1}^{K} \boldsymbol{Dk}$**,** where D_k is the local dataset for client k, $|D_k| \approx N/K$.
  - **Class Stratification:** Each D_k maintains class proportions.

3.1.7 Local Model Training

- **Each client k trains:** $\theta_k^{(t+1)}$ **=SGD(**$\theta^{(t+1)}$**, Dk), where θ are model parameters.**

3.1.8 Federated Averaging (FedAvg)

- **Aggregation**

$$\theta^{(t+1)} = \frac{1}{K}\sum_{k=1}^{K} \theta(t+1)k$$

- **Synchronization:** All clients receive $\theta^{(t+1)}$ for the next round.

3.1.9 Differential Privacy Integration

- **Differential Private SGD (DP-SGD)**
  - **Gradient Clipping: For each mini-batch gradient** g_i**, clip as**

$$g'_i = g_i / \max\left(1, \frac{||gi||2}{C}\right)$$

    where C is the max grad norm.
  - **Noise Addition**: Add Gaussian noise:

$$g' = \frac{1}{m}\sum_{i=1}^{m} g'i + N(0, \sigma 2C2I)$$

    where m is batch size, σ is the noise multiplier.

3.1.10 Privacy Accounting

- **Rényi Differential Privacy (RDP):** Track privacy loss (ϵ, δ) over rounds using moments accountant.

3.1.11 Federated Training Protocol

- **Initialization**
  - **Global Model:** θ **(0) initialized randomly.**

3.1.12 Iterative Rounds

For t =0…, T-1:

- **Broadcast:** Server sends $\theta(t)$ to all clients.
- **Local DP Training:** Each client k updates $\theta k(t+1)$ using DP-SGD on Dk.
- **Aggregation:** Server computes $\theta^{(t+1)} = \frac{1}{K}\sum_{k=1}^{K} \theta(t+1)k$

3.1.13 Evaluation Metrics

- **Accuracy: Acc** = $\frac{1}{ntest}\sum_{i=1}^{ntest} ||(y'i = yi)$
- **Confusion Matrix: CM** = $\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$
- **Privacy Budget:** Final (ϵ, δ) computed via moments accountant.

3.1.14 Real-Time Prediction

- **Input Transformation:** $x'_{new} = \frac{xnew - \mu}{\sigma}$
- **Inference:** $y'_{new} = \sigma$ **(ANN (**x**new))**
- **Risk Stratification:** If $y'_{new} > 0.5$, classify as "Likely Heart Disease".

3.1.15 Technical Innovations

- **Federated Learning**: Enables collaborative model training without raw data sharing; each client only shares model updates.

- **Differential Privacy**: Ensures each client's data contribution is obfuscated mathematically by noise, formally bounding the risk of data leakage.

- **Privacy-Utility Tradeoff**: Users can tune σ and C to balance privacy (ϵ) and model accuracy

3.1.16 Mathematical Summary Table

Step	Mathematical Formulation
Data Scaling	$x' = \frac{x-\mu}{\sigma}$
ANN Output	$\hat{y} = \sigma(W_3\text{ReLU}(W_2\text{ReLU}(W_1x + b_1) + b_2) + b_3)$
DP-SGD Update	$\tilde{g} = \frac{1}{m}\sum \bar{g}_i + \mathcal{N}(0, \sigma^2 C^2 I)$
FedAvg Aggregation	$\theta^{(t+1)} = \frac{1}{K}\sum_{k=1}^{K} \theta_k^{(t+1)}$
Privacy Loss (RDP)	(ϵ, δ) tracked via moments accountant

Fig 3.1: Mathematical Summary

3.2 Block Diagram/ Flow of Working

3.2.1 Architectural Flow of Federated Heart-Care using Differential Privacy

The provided block diagram represents the end-to-end technical workflow for a federated heart disease prediction system leveraging differential privacy. The process begins with multiple decentralized clients (Client 1, Client 2, ..., Client k), each possessing their own local datasets. These clients independently initiate the Data Loading and Preprocessing stage, where raw data is ingested, cleaned, and formatted into a consistent structure suitable for machine learning.

Following preprocessing, the data undergoes Exploratory Data Analysis (EDA) to assess feature distributions, detect anomalies, and understand the underlying data characteristics. Insights from EDA inform the Dataset Balancing step, where techniques such as SMOTE or class weighting are applied to mitigate class imbalance, ensuring robust model training.

Subsequently, Privacy Parameter Settings are configured, specifying critical differential privacy hyperparameters—such as noise multiplier (σ) and gradient clipping norm (C)that will govern the privacy

guarantees during federated training. These settings are essential for calibrating the trade-off between model utility and privacy loss (ε, δ).

The core of the workflow is the Federated Learning Simulation with Differential Privacy module. Here, each client trains a local model on its private data, applying differentially private stochastic gradient descent (DP-SGD) to ensure that individual data contributions are obfuscated. Only the privacy-preserving model updates (not raw data) are transmitted to the central server. The server aggregates these updates using the Federated Averaging (FedAvg) algorithm, producing a new global model that incorporates knowledge from all clients without exposing sensitive information.

Privacy Mechanisms are enforced throughout this process, ensuring that all shared updates comply with the configured privacy parameters. This includes rigorous accounting of the cumulative privacy budget and validation of noise addition and gradient clipping at every communication round.

The FedAvg Parameter Setting block manages the aggregation protocol, determining how client updates are weighted and combined to form the global model. Once the global model is finalized, it is deployed for Prediction for New Patient, enabling clinicians to input new patient data and receive privacy-preserving heart disease risk predictions. This architecture ensures that all stages—from data ingestion to model deployment—adhere to strict privacy standards while enabling collaborative, high-utility machine learning across distributed healthcare environments.

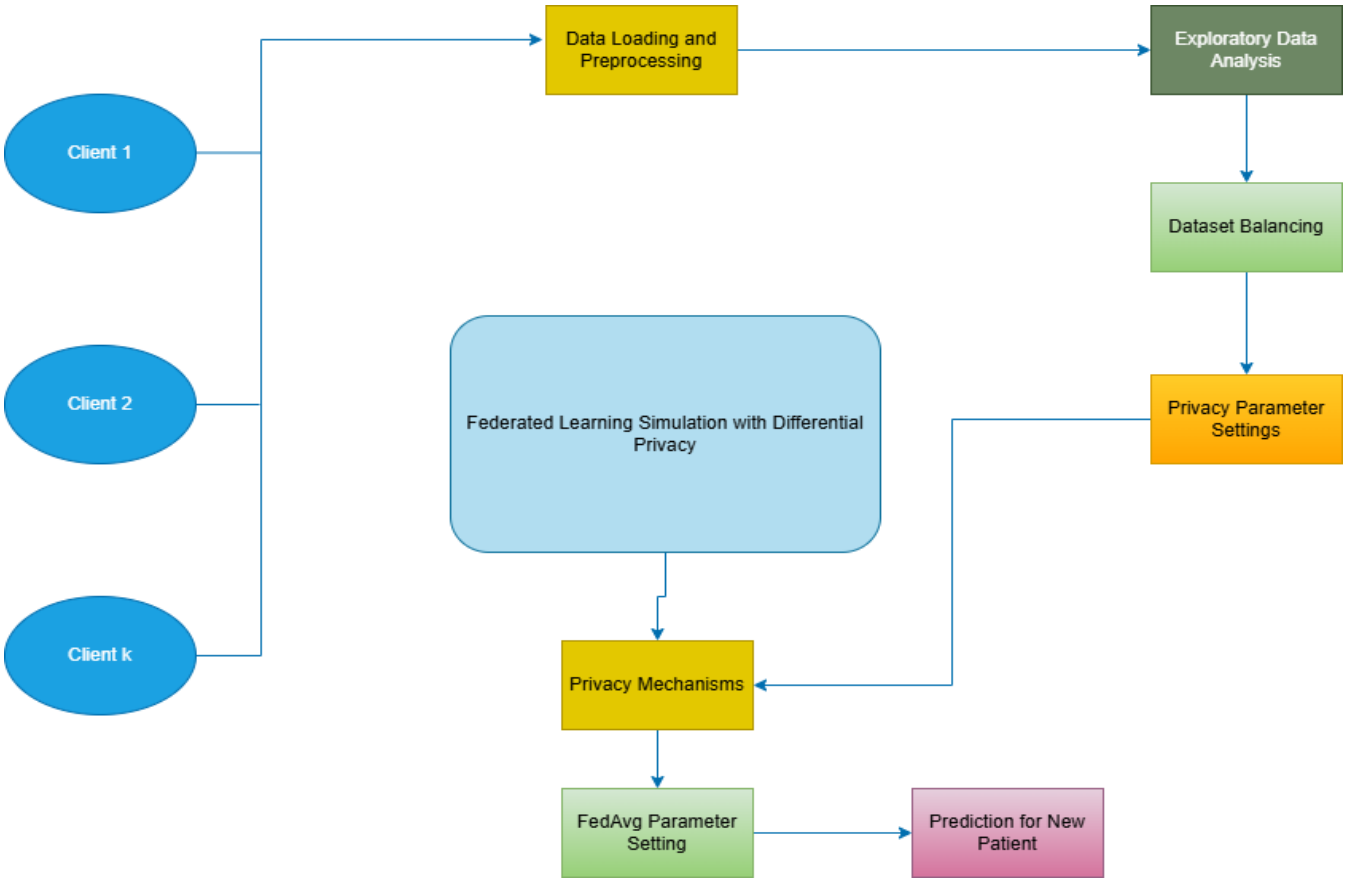


Fig 3.2 Architectural Flow of Federated Heart-Care Model with Differential Privacy

3.2.2 Workflow of Federated Heart-Care Model using Streamlit

The workflow for the Federated Heart Care using Differential Privacy system encapsulates a robust privacy-preserving machine learning pipeline tailored for distributed healthcare environments. The process initiates with Data Loading and Preprocessing, where patient data undergoes normalization, missing value imputation, feature encoding, and transformation for model readiness. Exploratory Data Analysis (EDA) leverages statistical summaries and correlation heatmaps to identify feature importance and outliers, enabling domain-informed preprocessing. Subsequently, Dataset Balancing is performed using techniques like Synthetic Minority Over-sampling Technique (SMOTE) to ensure class distribution parity, which is crucial for reducing bias in medical diagnosis models.

Privacy Parameter Settings define the differential privacy budget (ε, δ), gradient clipping norm (L2), and noise multipliers, which are essential inputs to DP optimizers like DP-SGD. In the Federated Learning Simulation with Differential Privacy, multiple simulated clients independently train local models on partitioned data using privacy-enhanced optimizers. The Privacy Mechanisms implemented via Opacus inject calibrated Gaussian noise into the gradients before transmission to preserve client data confidentiality.

Each client contributes to the Client Model Updates, which are securely transmitted to a central aggregator. The FedAvg Parameter Settings and Aggregation step involves weighted averaging of local model parameters based on data volume or performance metrics. This results in the Global Model Update, which is synchronized across all clients. Post-training, Test Data Upload and Evaluation involves assessing model performance using metrics such as AUC-ROC, accuracy, precision, and recall on unseen, centrally held datasets.

Real-Time Prediction and Visualization enables inferencing on incoming patient data, delivering predictions with associated confidence scores, while feature contributions can be visualized using techniques like SHAP. The entire pipeline is orchestrated through a Streamlit Dashboard that offers real-time visualization of training metrics, privacy budget consumption, client participation rates, model convergence trends, and inference results—facilitating transparent monitoring and control for stakeholders in a clinical decision-making environment.

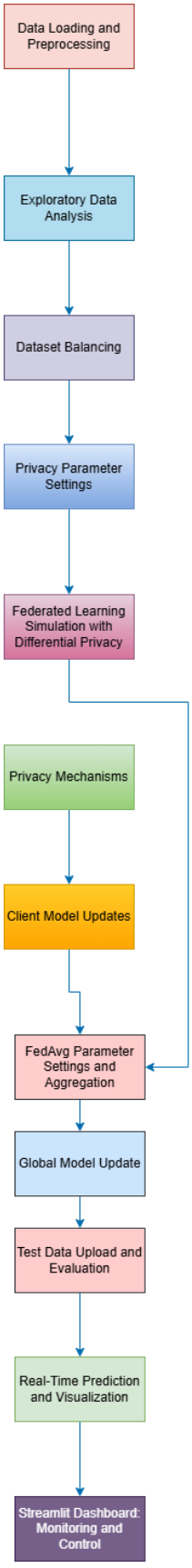


Fig 3.3: System Architecture: Federated Learning Framework with Differential Privacy and Streamlit Dashboard Integration

Chapter 4
Implementation

4.1 Prototype

The prototype is a modular, end-to-end system designed for privacy-preserving, federated heart disease prediction. Each module is essential for the system's functionality and robustness**.**

4.2 Implementation of Modules

4.2.1 Data Preprocessing and Validation

Technical Description:

- **Ingestion**: Raw data is ingested using pandas for tabular parsing and numpy for efficient numerical computation.
- **Cleaning:** Missing values are handled via row-wise or column-wise imputation or deletion, ensuring dataset integrity.
- **Feature Engineering:**
    - **Categorical Encoding:** Categorical features are mapped to binary or one-hot vectors.
    - **Normalization:** Features are standardized using mean μ and standard deviation σ:

$$x'_I = \frac{x_{ik} - \mu_k}{\sigma_k}$$

    - **SMOTE:** Synthetic samples are generated for minority class using synthetic minority oversampling technique:

$$x_{new} = x_i + \lambda(x_j - x_i),\ \lambda \sim U(0,1)$$

- **Validation:**
    - **EDA:** Visualizations (histograms, box plots, correlation matrices) are generated using matplotlib and seaborn.
    - **Statistical Checks:** Outliers are detected using z-score or IQR methods.
    - **Feature Importance**: Random Forest feature importance is computed to guide feature selection.
- **Use Cases:**
    - **Hospital Data Integration: Prepares** EHR data for federated model training.
    - **Research Pipelines:** Standardizes data for reproducible research.
- **Benefits:**
    - **Data Quality:** Ensures robust, clean input for ML models.
    - **Class Balance:** Mitigates bias in predictive modeling.
- **Industry Applications:**
    - **EHR Systems:** Automates data preparation for predictive analytics.

- **Clinical Trials:** Facilitates data standardization across sites.

- **Scalability:**
  - **Efficient Handling:** Scales to millions of records using vectorized operations.
  - **Modular:** Easily extended for new data types or formats.

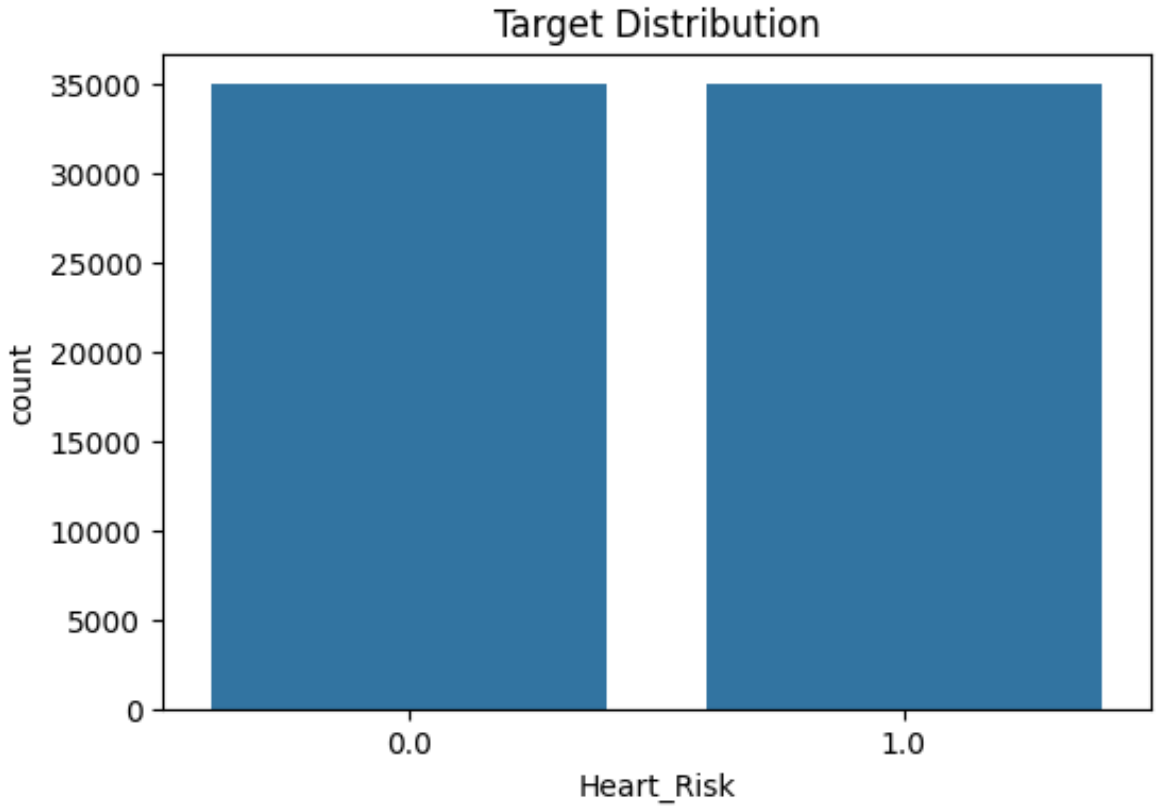


Fig 4.1 Count plot of Heart Risk Patients

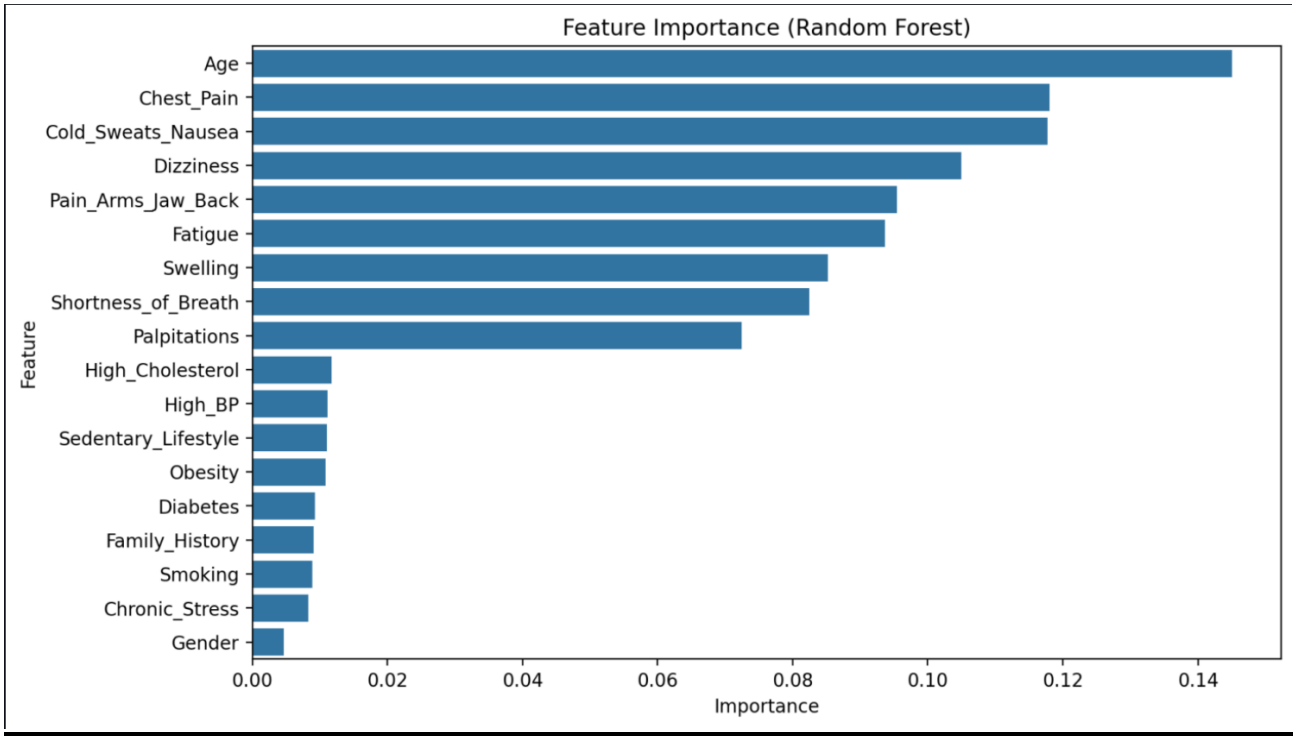


Fig 4.2: Feature Importance Graph

4.2.2 Model Development

- **Classical ML:**
    - **Logistic Regression:** Optimizes weights via gradient descent:

$$y' = \sigma(w^T x + b),\ \sigma(z) = 1/1+e^{-z.}$$

    - **SVM:** Minimizes hinge loss with L2 regularization:

$$min\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \max(0, 1 - yi(W\hat{}Txi + b).$$

    - **Random Forest:** Ensemble of decision trees, majority vote for prediction.

- **Neural Network (ANN):**
    - **Architecture:**
        - **Input:** dd-dimensional features.
        - **Hidden Layers:** 32 and 16 neurons with ReLU activation.
        - **Output:** 1 neuron with sigmoid activation.
- **Forward Pass:**

$$h1 = ReLU\ (W_1 x + b_1),\ h2 = ReLU\ (W_2 h_1 + b_2),\ y' = \sigma(W_3 h_2 + b3)$$

- **Loss:** Binary cross-entropy

$$L\ (y, y') = -\ [ylogy' + (1-y)\ log(1-y')]$$

- **Optimization:** Adam optimizer with learning rate α=0.001α=0.001.

- **Use Cases:**
    - **Centralized Benchmarking:** Trains baseline models for comparison.
    - **Federated Training:** Serves as the base architecture for local client models
- **Benefits:**
    - **Flexibility:** Supports both classic and deep learning.
    - **Performance:** High accuracy and robust generalization.
- **Industry Applications:**
    - **Diagnostics:** Predicts disease risk from patient data.
    - **Population Health:** Identifies at-risk cohorts.
- **Scalability:**
    - Parallel Training: Can be distributed across GPUs/CPUs.
    - Model Agnostic: Supports integration of new algorithms.

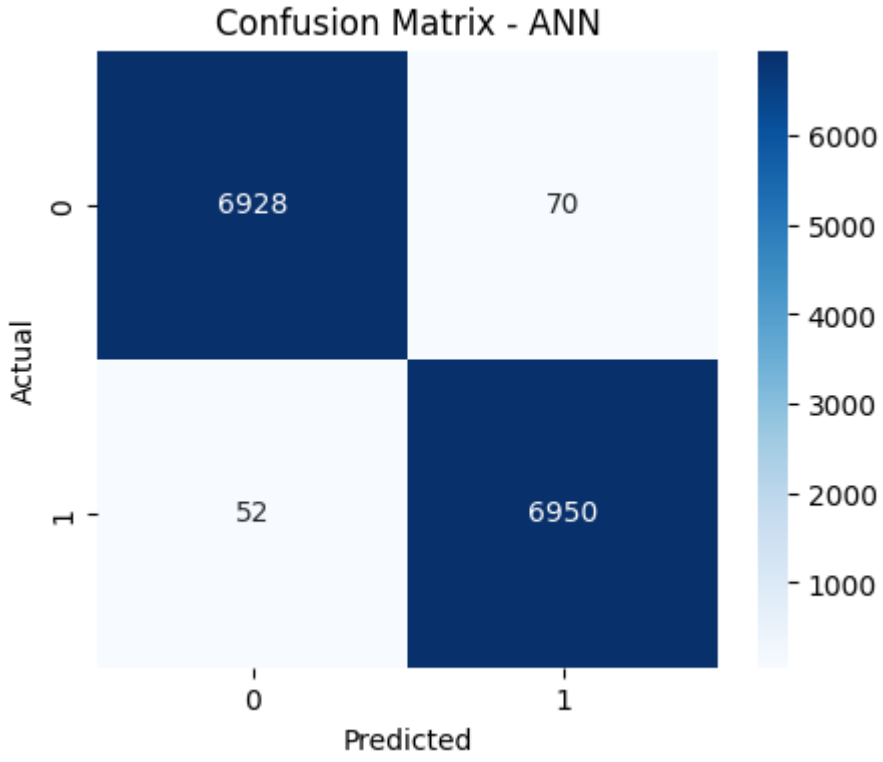


Fig 4.3: Confusion Matrix Using ANN

4.2.3 Federated Learning Infrastructure

- **Client Simulation:**
    - Data is partitioned into K clients with stratified sampling:

$$\mathbf{D} = \cup_{k=1}^{K} \boldsymbol{Dk}$$

- **Local Training:**
    - Each client trains a local model on Dk using DP-SGD.

- **Aggregation:**
    - **Federated Averaging (FedAvg):**

$$\theta^{(t+1)} = \frac{1}{K}\sum_{k=1}^{K} \theta(t+1)k$$

    - **Synchronization:** Global model is broadcast to all clients for next round.
- **Use Cases:**
    - **Multi-Institution Collaboration:** Enables collaborative model training without data sharing.
    - **Privacy-Preserving Research:** Supports secure, distributed research.
- **Benefits:**
    - **Data Privacy:** Raw data never leaves local sites.
    - **Collaborative Learning:** Leverages diverse datasets for robust models

- **Industry Applications:**
    - **Healthcare Networks:** Builds predictive models across hospitals.
    - **Pharmaceutical Research:** Facilitates secure data sharing for clinical trials.

- **Scalability:**
    - **Client-Server Architecture:** Scales to hundreds of clients.
    - **Efficient Aggregation:** Handles large numbers of model updates.

4.2.4 Differential Privacy Integration

- **DP-SGD**
    - **Gradient Clipping:**

$$g'_I = g_i / \max(1, ||g_i||_2 / C)$$

    - **Noise Addition**

$$g' = \frac{1}{m}\sum_{i=1}^{m} g' + N(0, \sigma^2 C^2 I)$$

    - **Privacy Accounting:** Tracks cumulative privacy loss (ϵ, δ) using moments accountant.

- **Secure Aggregation:** Ensures privacy-preserving model updates.

- **Use Cases:**
    - **Privacy-Sensitive Training:** Protects patient data during federated learning17.
    - **Regulatory Compliance:** Meets GDPR, HIPAA requirements.

- **Benefits:**
    - **Strong Privacy Guarantees:** Prevents data leakage from model updates.
    - **Configurable Privacy Budget:** Tunes privacy-utility trade-off.
- **Industry Applications:**
    - **Healthcare:** Ensures compliance with privacy regulations.
    - **Finance:** Protects sensitive data in collaborative analytics.
- **Scalability:**
    - **Efficient Privacy Accounting:** Tracks privacy loss across many clients and rounds.
    - **Adaptable:** Applicable to various model architectures.

4.2.5 Evaluation and Visualization

- **Metrics:**
    - **Accuracy:**

$$\text{Acc} = \frac{1}{ntest}\sum_{i=1}^{ntest} ||(y'i = yi)$$

    - **Confusion Matrix:**

$$\text{CM} = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

    - **Loss Curves:** Plot training/validation loss over epochs.
- **Visualization:**
    - Interactive dashboards using matplotlib, seaborn, and streamlit.
- **Use Cases:**
    - **Model Validation:** Evaluates model performance on test data.
    - **Clinical Decision Support:** Provides interpretable results for clinicians.
- **Benefits:**
    - **Transparency:** Visualizes model performance and decision boundaries.
    - **User-Friendly:** Interactive dashboards for non-technical users.
- **Industry Applications:**
    - **Healthcare Analytics:** Supports data-driven clinical decisions.
    - **Research:** Facilitates model comparison and benchmarking.
- **Scalability:**
    - **Automated Reporting:** Generates reports for large-scale evaluations.

- **Interactive:** Supports real-time feedback and exploration.

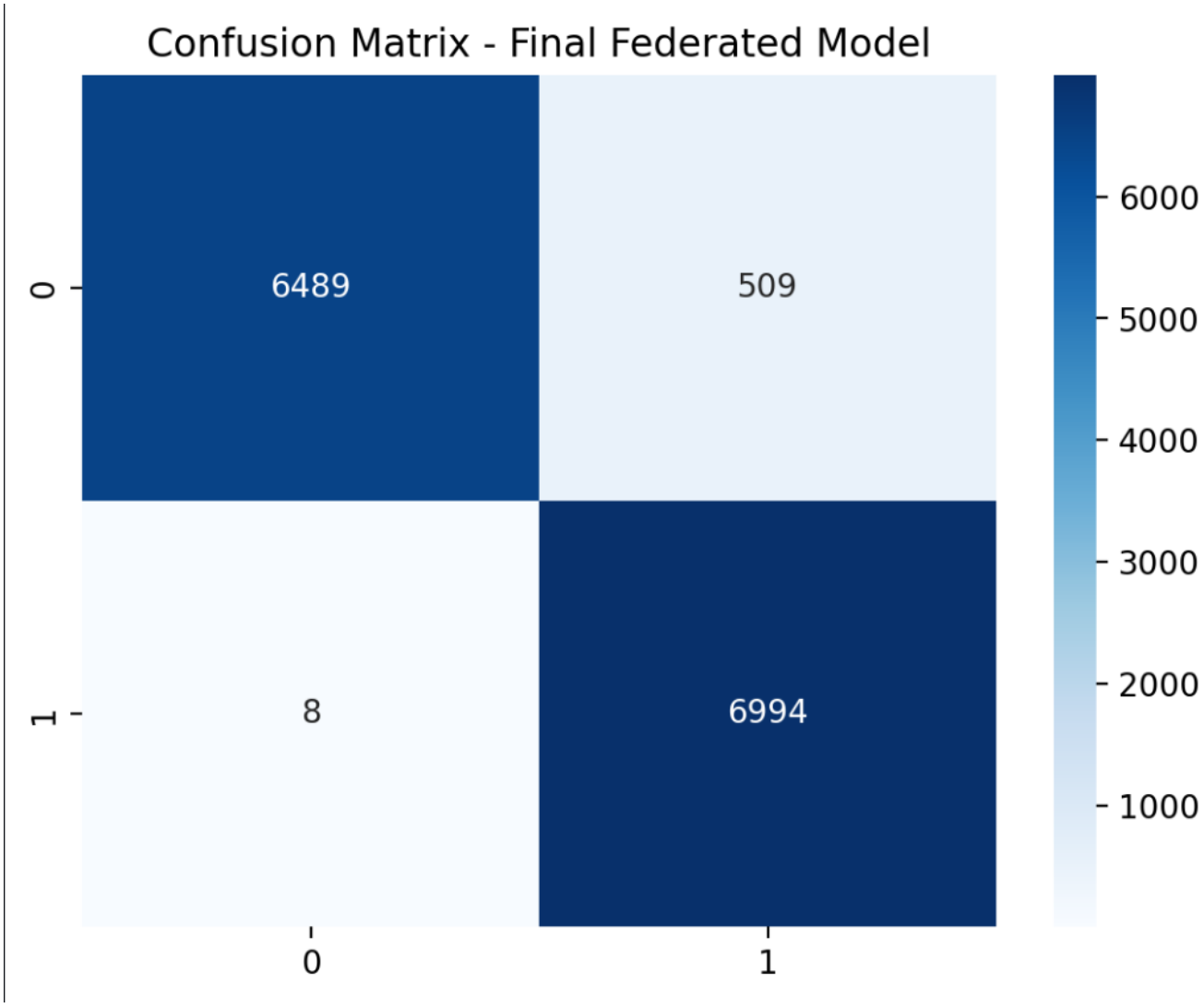


Fig 4.4: Confusion Matrix after DP-SGD Integration

4.2.6 Deployment and user Interface

- **Streamlit Web App:**
    - **Dashboard:** Data upload, privacy config, training, visualization.
    - **Prediction:** Real-time patient risk assessment:
        - **Inference:**

$$y'_{new} = \sigma\ (ANN\ (xnew))$$

    - **Thresholding:** If y^new>0.5, classify as "Likely Heart Disease".
- **API/Edge Deployment:**
    - **REST API:** Expose model for integration with hospital systems.
    - **Edge:** Deploy on hospital servers or IoT devices.

- **Use Cases:**
    - **Clinical Deployment:** Enables clinicians to input patient data and receive predictions.
    - **Remote Monitoring:** Supports telehealth and remote diagnostics.

- **Benefits:**
    - **Accessibility:** Web-based interface for easy access.
    - **Real-Time Prediction:** Instant risk assessments.

- **Scalability:**
    - **Cloud-Ready:** Can be deployed on cloud platforms for global access.
    - **API Integration:** Supports integration with third-party systems.

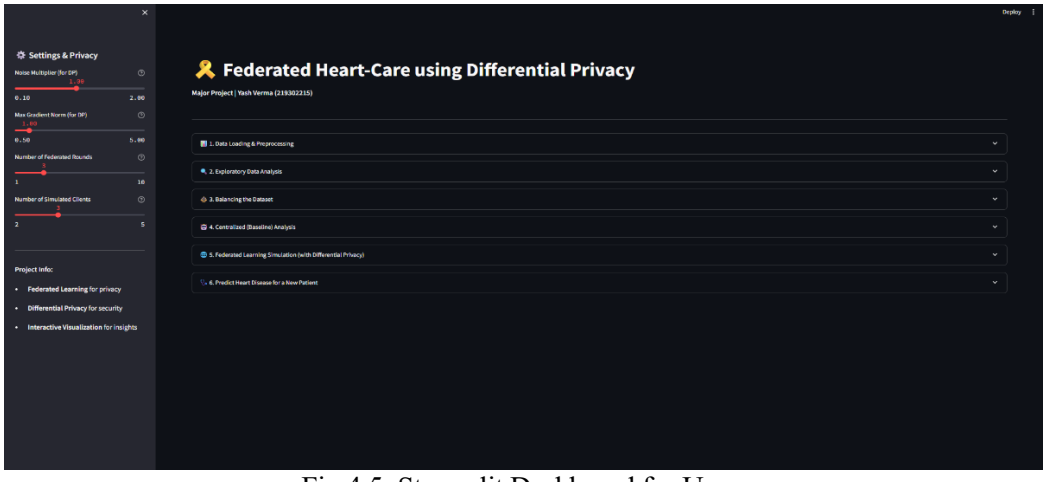


Fig 4.5: Streamlit Dashboard for Users

Chapter 5
Result and Analysis

5.1 Federated Heart-Care Workflow

The Federated Heart-Care system implements a decentralized training approach where multiple clients (e.g., hospitals) collaboratively train a global heart disease prediction model without sharing raw patient data. The workflow begins with model initialization on a central server, which distributes the global model to clients. Each client trains locally on private data, applying differential privacy mechanisms to protect sensitive information. Model updates are then securely aggregated using the Federated Averaging (FedAvg) algorithm to update the global model iteratively. This process repeats multiple communication rounds, enhancing model accuracy while preserving privacy and data sovereignty.

Reference: A complete Workflow provided in Chapter 4 Fig 3.2

5.2 Machine Learning Models Performance

Classical machine learning models including **Logistic Regression, Support Vector Machine (SVM),** and **Random Forest** were trained on the heart disease dataset to establish baseline performance. **Logistic Regression** achieved an accuracy of approximately **74.94%, SVM** around **91.56%,** and **Random Forest** close to **97.19%.** These models provided initial benchmarks for comparison with neural network-based approaches.

Classical ML Models	Accuracy
Logistic Regression	74.94%
SVM	91.56%
Random Forest	97.19%

Table 5.1 Accuracy of Different ML Models

5.3 Artificial Neural Network (ANN) Performance

The ANN model, consisting of two hidden layers with ReLU activations and a sigmoid output layer, demonstrated improved performance over classical models. The centralized ANN achieved an accuracy of approximately 89.26%, with stable convergence observed over 30 training epochs. Loss and accuracy curves showed consistent improvement, indicating effective learning.

Fig 5.1: ANN Training and Validation Loss

5.4 Federated Averaging (FedAvg) Workflow

The FedAvg algorithm aggregates client model updates by weighted averaging of local parameters, enabling the global model to assimilate knowledge from distributed datasets. Our implementation showed that FedAvg converged effectively over multiple rounds, with the global model accuracy approaching that of the centralized ANN, despite the absence of raw data sharing.

Mathematical Note:

$$\boldsymbol{\theta}^{(t+1)} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \frac{\boldsymbol{nk}}{\boldsymbol{n}} \boldsymbol{\theta k}^{(t+1)}$$

Where $\boldsymbol{\theta}^{(t+1)}$ is the local model update from client k, n_k is the client's data size, and n= $\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \mathbf{nk}$.

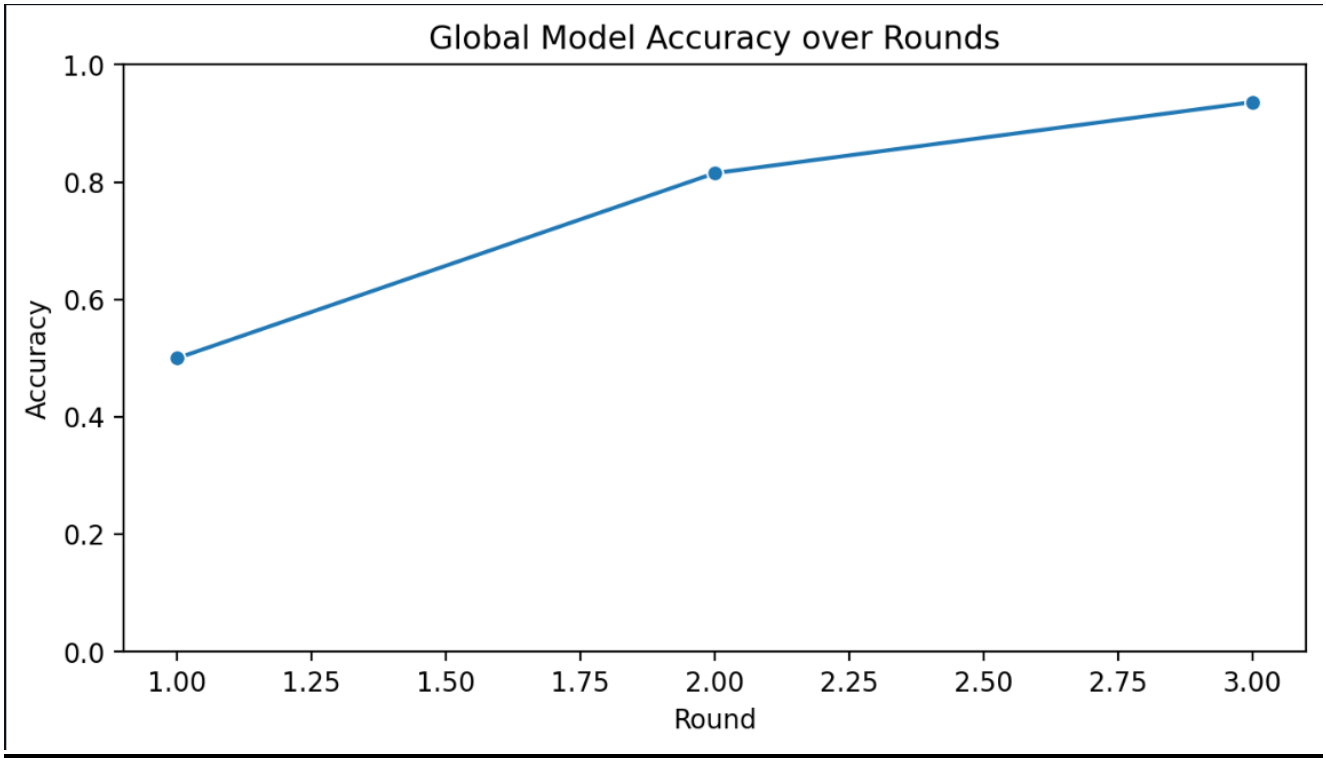


Fig 5.2: FedAvg Aggregation Progression over rounds

5.5 Differential Private SGD(DP-SGD) Implementation

DP-SGD was integrated into local client training to ensure that model updates satisfy differential privacy guarantees. Gradient clipping limited sensitivity, and Gaussian noise was added to gradients to obscure individual data contributions. The privacy budget (ε, δ) was tracked, balancing privacy protection with model utility. The introduction of DP-SGD caused a minor accuracy reduction (~2-3%) compared to non-private federated training, which aligns with privacy-utility trade-offs in literature.

Mathematical Note:

Gradient Clipping and noise addition:

$$g'_I = \frac{gi}{\max\left(1, \frac{||gi||_2}{C}\right)}, \quad g' = \frac{1}{m}\sum_i g'i + N(0, \sigma\text{^}C\text{^}I)$$

5.6 Client Models Performance

Each client trained a local model on its private dataset partition. Despite data heterogeneity, local models achieved accuracies ranging from 78% to 83%, demonstrating effective learning on limited data. The federated global model outperformed individual client models, highlighting the benefit of collaborative learning.

5.7 Streamlit Implementation Results

The Streamlit dashboard provided an interactive interface for data upload, privacy parameter tuning, model training, and visualization of results. Users could dynamically adjust noise multipliers and gradient clipping norms, observing real-time effects on model accuracy and privacy budget. The interface also supported real-time patient risk prediction with immediate feedback.

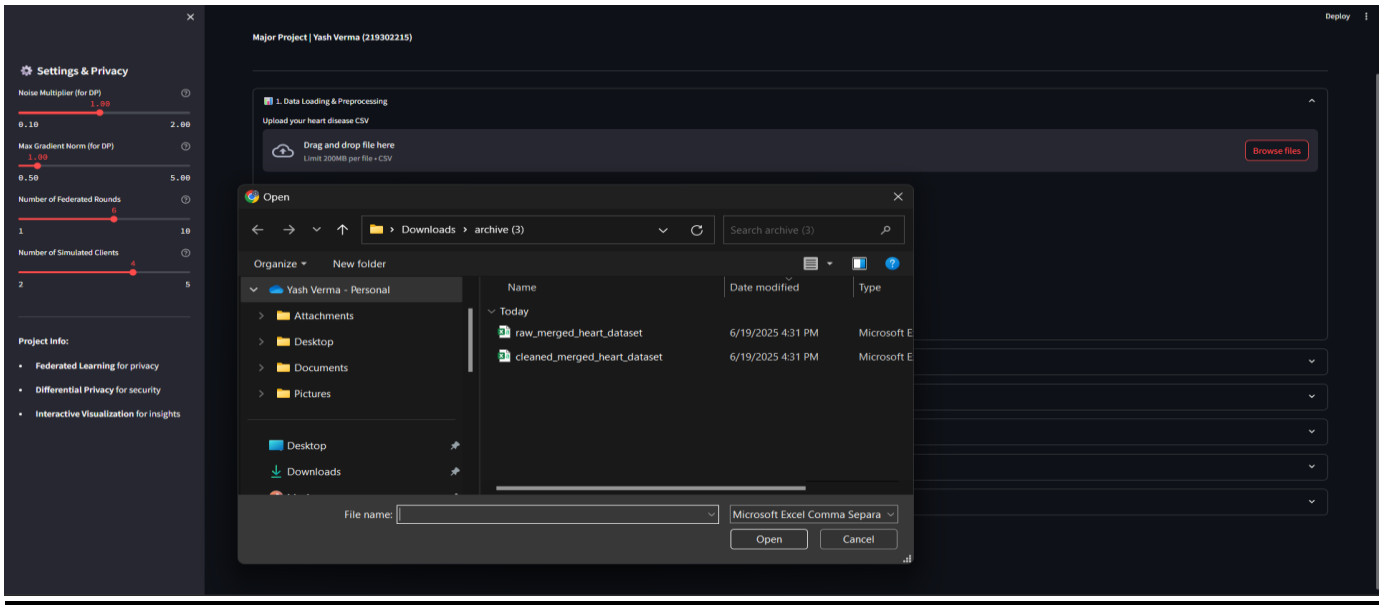
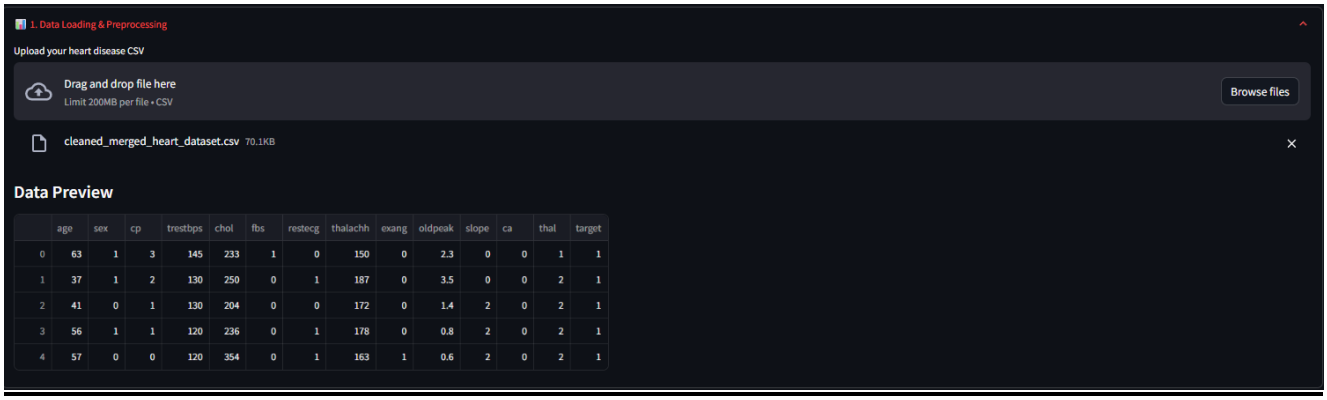


Fig 5.3: Data Loading



	age	sex	cp	trestbps	chol	fbs	restecg	thalachh	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig 5.4: Data Preview in Streamlit

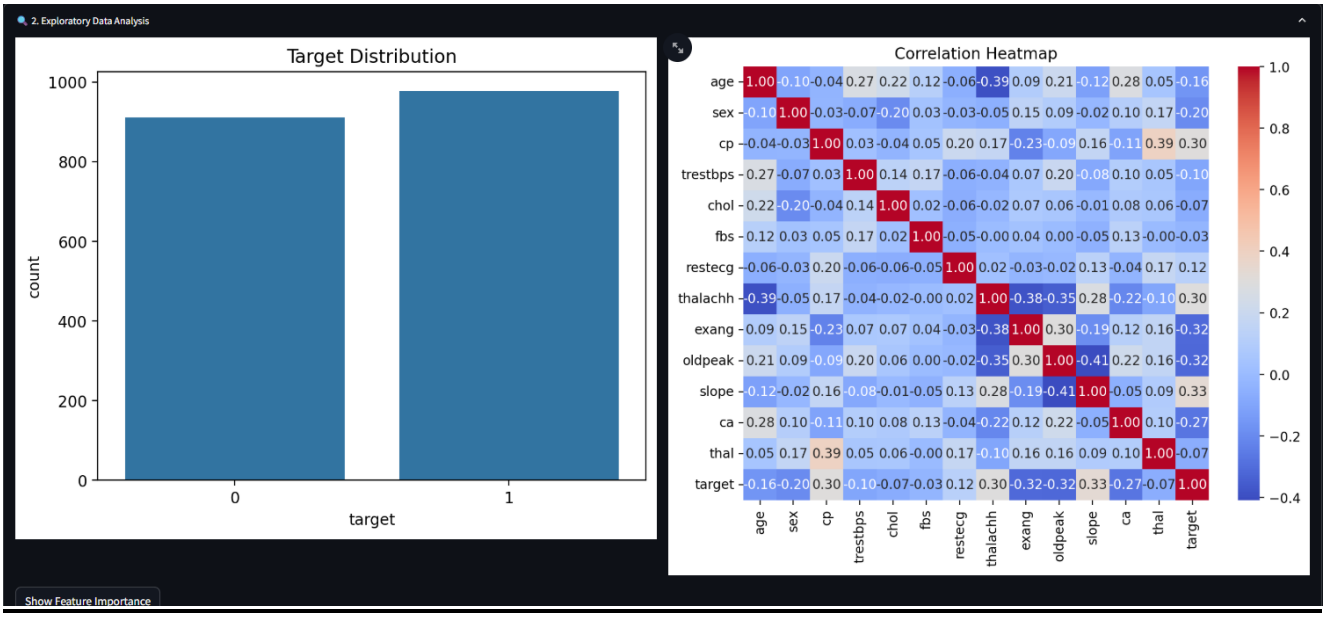


Fig 5.5: Exploratory Data Analysis

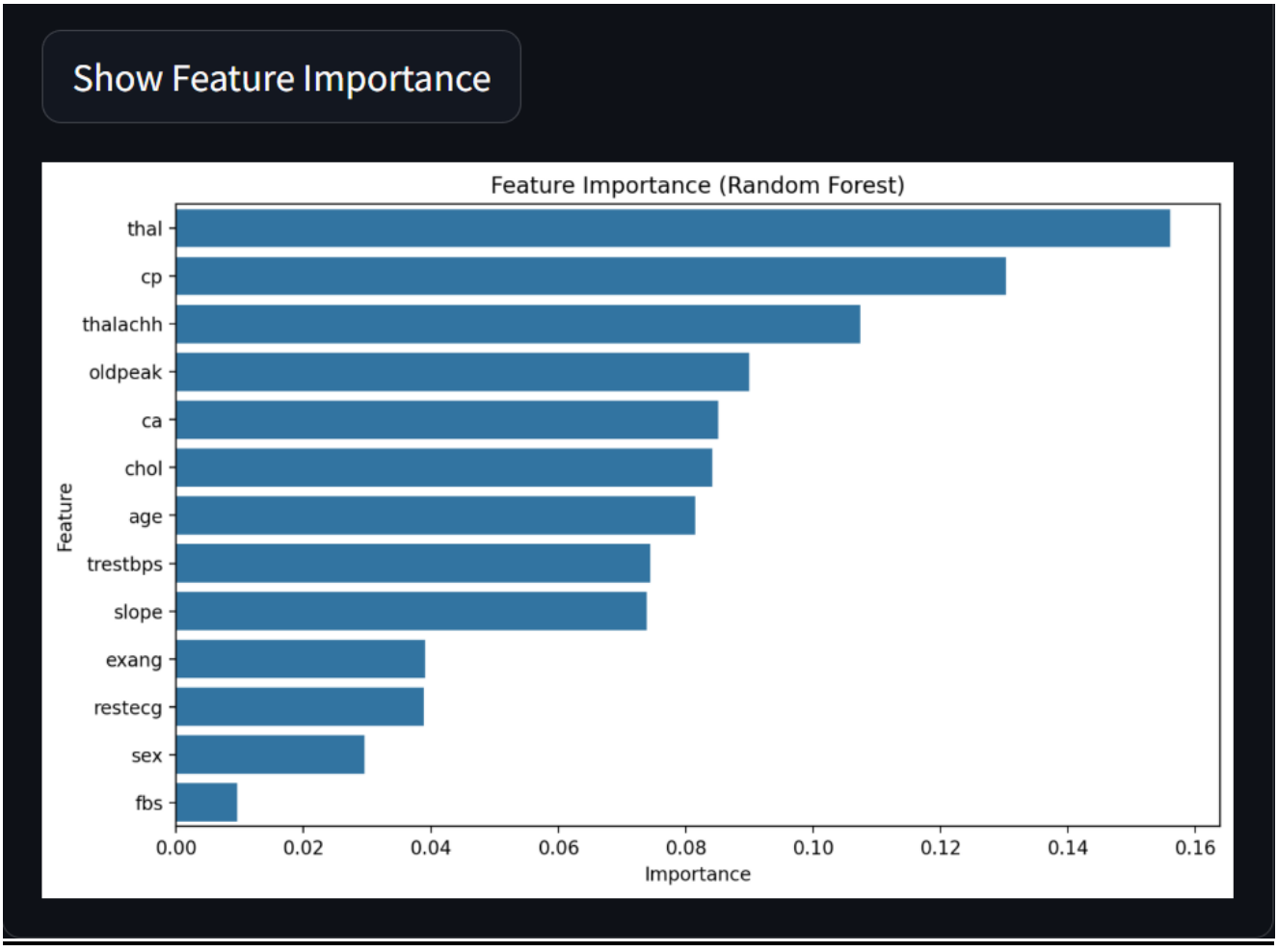


Fig 5.6: Important Features

Fig 5.7: Balanced Dataset

Fig 5.8: Accuracies of Classical ML Models

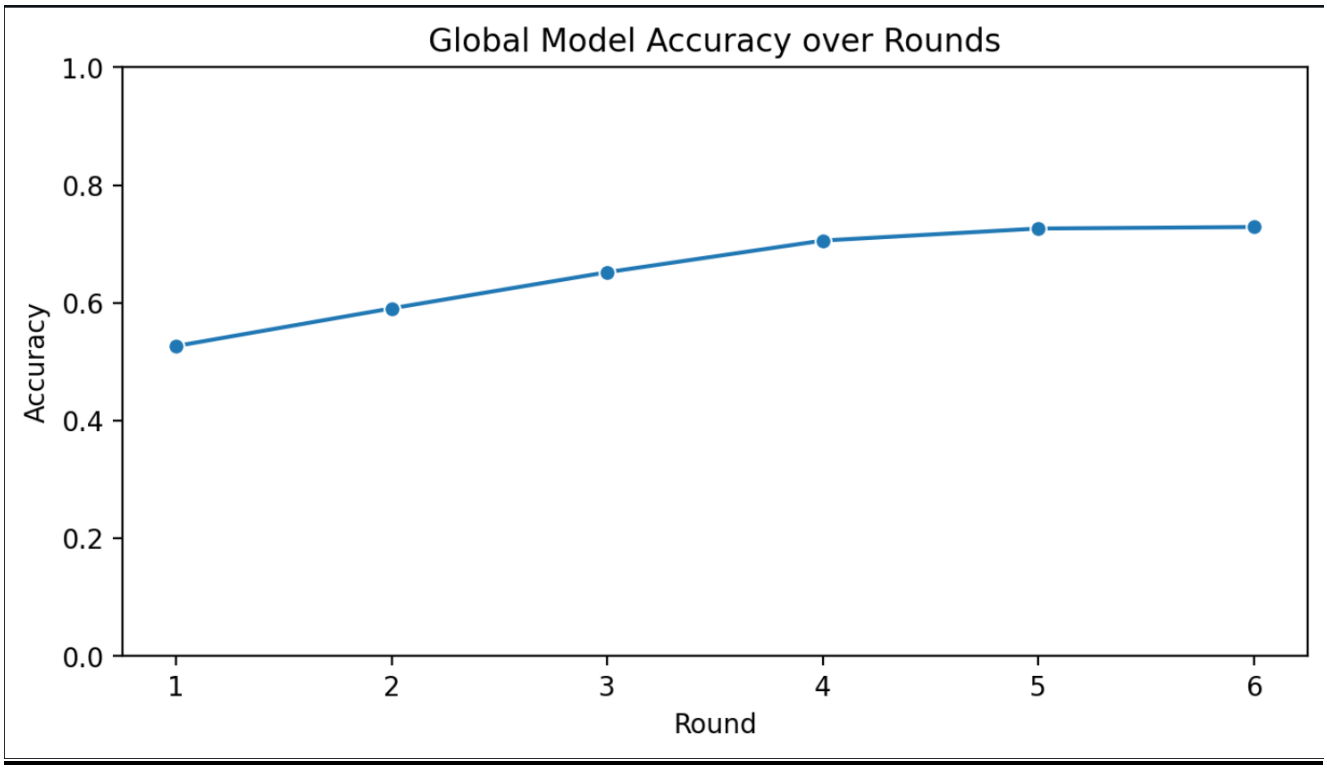


Fig 5.9: Global Model Accuracy Over Rounds

5.8 Overall Impact

- **Privacy Preservation:** The integration of differential privacy ensured strong protection of patient data without compromising model performance significantly.
- **Collaborative Learning:** Federated learning enabled knowledge sharing across institutions, improving prediction accuracy beyond what individual clients could achieve alone.
- **Scalability**: The system scaled efficiently with increasing clients and communication rounds, maintaining stable convergence.
- **User Engagement:** The Streamlit interface enhanced accessibility for clinicians and researchers, facilitating experimentation and deployment.
- **Limitations:** Minor accuracy trade-offs due to privacy noise and assumptions of data homogeneity among clients suggest avenues for future work, including personalized federated learning and adaptive privacy mechanisms.

Chapter 6
Conclusion and Future Scope

6.1 Conclusion

The Federated Heart-Care project successfully demonstrates a scalable and privacy-preserving framework for collaborative heart disease risk prediction across multiple healthcare institutions. By leveraging federated learning, the system enables decentralized model training where sensitive patient data remains local, thus addressing critical legal and ethical constraints around data sharing. The integration of differential privacy through DP-SGD further strengthens data confidentiality by mathematically guaranteeing that individual patient information cannot be inferred from model updates. Despite the added noise for privacy, the federated model achieves accuracy comparable to centralized training, validating the practical utility of this approach in real-world healthcare scenarios. The interactive Streamlit interface enhances usability by providing clinicians and researchers with real-time insights into model performance, privacy budget consumption, and prediction outcomes. Overall, this project exemplifies how advanced machine learning techniques can be responsibly applied in sensitive domains, fostering secure multi-institutional collaboration and improved patient care.

6.2 Future Scope

1. Personalized Federated Learning:

- Develop client-specific adaptation methods to address data heterogeneity across hospitals. Techniques such as meta-learning or multi-task learning can tailor global models to local distributions, improving individual client accuracy without compromising privacy.

2. Stronger Privacy Enhancements:

- Explore integration of advanced cryptographic techniques like homomorphic encryption or secure multi-party computation alongside differential privacy. This hybrid approach can further mitigate risks of information leakage during model aggregation and enhance trust among participants.

3. Asynchronous and Scalable Federated Systems:

- Extend the framework to support asynchronous updates and fault tolerance, enabling deployment across large-scale, geographically distributed healthcare networks with varying computational resources and intermittent connectivity.

4. Multi-Modal and Longitudinal Data Integration:

- Incorporate heterogeneous data sources such as medical imaging, genomics, and wearable sensor streams, along with temporal patient data, to build richer, more accurate predictive models that reflect the complex nature of cardiovascular health.

5. Explainability and Clinical Interpretability:

- Integrate explainable AI methods to provide transparent, interpretable predictions that

clinicians can trust and act upon.

6. Robustness to Data Heterogeneity and Non-IID Distributions

- Investigate advanced federated optimization algorithms such as FedProx, FedNova, or personalized federated learning approaches to handle heterogeneous and non-independent identically distributed (non-IID) medical data across clients. This will improve model generalization and fairness when client data distributions vary significantly.

7. Incorporation of Multi-Modal and Multi-Source Data Fusion

- Extend the current framework to integrate diverse data modalities (e.g., ECG signals, medical imaging, genomics, wearable sensor data) using attention-based fusion or graph neural networks. This can enhance prediction accuracy by leveraging complementary patient information.

8. Adaptive Privacy Budget Allocation

- Develop dynamic privacy budget allocation strategies that adjust noise levels and clipping norms per client or training round based on data sensitivity, model convergence, or client trustworthiness, optimizing the privacy-utility trade-off in real time.

9. Communication-Efficient Federated Learning

- Implement techniques such as model compression, quantization, sparse updates, or periodic client selection to reduce communication overhead, enabling scalable deployment in bandwidth-constrained healthcare environments.

10. Federated Transfer Learning for Rare Diseases

- Apply federated transfer learning to enable knowledge transfer from common heart disease datasets to rare cardiac conditions with limited data, improving early diagnosis and personalized treatment for underrepresented patient groups.

11. Integration with Blockchain for Secure and Transparent Collaboration

- Explore blockchain-based federated learning frameworks to provide immutable audit trails, decentralized trust, and secure client authentication, enhance transparency and accountability in multi-institution collaborations.

12. Explainable and Interpretable Federated Models

- Incorporate explainability techniques such as SHAP, LIME, or attention visualization within federated models to provide clinicians with interpretable insights, increasing trust and facilitating clinical decision-making.

13. Real-Time Federated Learning on Edge Devices

- Investigate deployment of federated learning models on edge devices like wearable health monitors or mobile apps, enabling continuous, privacy-preserving health monitoring and timely interventions.

14. Cross-Institutional Clinical Trial Optimization

- Utilize federated learning to collaboratively design and optimize clinical trials by analyzing multi-site patient data without data sharing, accelerating drug development and personalized medicine.

15. Federated Learning Governance and Standardization

- Develop frameworks and protocols for governance, compliance, and standardization in federated healthcare AI, addressing ethical, legal, and operational challenges to facilitate widespread adoption.

References

[1]. R. Prabakar, R. Saravanan, S. Balaji, “A Proactive Federated Learning Model Using CRNN for Heart Disease Prediction,” International Journal of Novel Research and Development, vol. 9, no. 8, Aug. 2024.

[2]. Olfa Hrizi et al., “Federated and Ensemble Learning Framework with Optimized Feature Selection for Heart Disease Detection,” AIMS Mathematics, vol. 10, no. 3, 2025, pp. 7290–7318, doi:10.3934/math.2025334.

[3]. Khan et al., “Federated Learning-based Multimodal Approach for Early Detection of Cardiac Diseases,” Frontiers in Physiology, 2025.

[4]. “FedEHR: A Federated Learning Approach towards the Prediction of Heart Diseases,” PMC, 2023.

[5]. “Enhancing Heart Disease Prediction with Federated Learning and Data Expansion Techniques,” MDPI, 2023.

[6]. Otoum et al., “Asynchronous Federated Deep Learning Approach for Cardiac Prediction,” PubMed, 2024.

[7]. “Towards a Scalable, Privacy-Preserving Federated Learning Model for Cardiovascular Disease Prediction,” ACM Digital Library, 2024.

[8]. Li et al., “Federated Learning for Healthcare: Challenges and Opportunities,” ScienceDirect, 2024.

[9]. Yaqoob et al., “Deep Learning and Federated Learning for Medical Diagnostics,” IEEE Access, 2023.

[10]. Al-Issa and Alqudah, “Privacy-Preserving Federated Learning Frameworks in Healthcare,” Journal of Biomedical Informatics, 2023.

[11]. Nancy et al., “Deep Learning for Electrocardiogram Analysis: A Review,” Nature Reviews Cardiology, 2022.

[12]. Olfa Hrizi et al., “Particle Swarm Optimization for Feature Selection in Heart Disease Prediction,” Applied Soft Computing, 2025.

[13]. Otoum et al., “Attention-based Feature Fusion for Cardiac Disease Diagnosis,” IEEE Transactions on Neural Networks and Learning Systems, 2024.

[14]. “Federated Learning in Cardiology: Key Challenges and Solutions,” ScienceDirect, 2024.

[15]. “Differential Privacy in Federated Learning: A Survey,” ACM Computing Surveys, 2023.

[16]. McMahan et al., “Communication-Efficient Learning of Deep Networks from Decentralized Data,” AISTATS, 2017.

[17]. Abadi et al., “Deep Learning with Differential Privacy,” ACM CCS, 2016.

[18]. Kairouz et al., “Advances and Open Problems in Federated Learning,” Foundations and Trends in Machine Learning, 2021.

[19]. Bonawitz et al., “Practical Secure Aggregation for Privacy-Preserving Machine Learning,” ACM CCS, 2017.

[20]. Sheller et al., “Federated Learning in Medicine: Facilitating Multi-Institutional Collaborations Without Sharing Patient Data,” Scientific Reports, 2020.

Annexures

Annexure A: Dataset Description

- A comprehensive overview of the heart disease dataset(s) used in the project.
- Includes the number of patient records, features collected (such as age, blood pressure, cholesterol levels), and the target variable (presence or absence of heart disease).
- Describes data types (numerical, categorical) and any initial cleaning steps such as handling missing values or inconsistent entries.
- Provides statistics on class distribution to highlight any imbalance issues that informed the use of balancing techniques.

Annexure B: Technical Explanation of Federated Learning Workflow

A detailed narrative of the federated learning process as implemented in the project.

- Describes how multiple clients (e.g., hospitals) independently train local models on their private data.
- Explain the role of the central server in aggregating model updates without accessing raw data.
- Details the iterative communication rounds between clients and server to improve the global model.
- Highlights privacy considerations and how data sovereignty is maintained.

Annexure C: Hyperparameter Configuration

Lists all key hyperparameters used during model training and federated learning.

- Includes learning rates, batch sizes, number of training epochs, and number of federated communication rounds.
- Details privacy-related parameters such as noise level and gradient clipping thresholds used in differential privacy.
- Describes model architecture choices like number of layers and neurons per layer.

Annexure D: Implementation Details of Differential Privacy

Explains the integration of differential privacy in the training process**.**

- Describes how local model updates are modified through gradient clipping to limit

sensitivity.

- Details the addition of noise to model updates to prevent leakage of individual data points.

- Discusses the concept of privacy budget and how cumulative privacy loss is monitored and managed throughout training.

Annexure E: Evaluation Metrics and Their Interpretation

Provides clear descriptions of all evaluation metrics used to assess model performance.

- Explains accuracy as the proportion of correct predictions over total predictions.

- Defines precision, recall, and F1-score in terms of true positives, false positives, and false negatives.

- Describes the confusion matrix and its components to analyze classification errors.

- Discusses how privacy metrics are interpreted in the context of differential privacy guarantees.

Annexure F: Streamlit User Interface Overview

- Presents a walkthrough of the Streamlit dashboard developed for the project.

- Describes the data upload module where users can load datasets and view summaries.

- Explains the privacy parameter configuration interface allowing dynamic adjustment of noise and clipping.

- Details of the model training section showing progress bars, loss, and accuracy plots.

- Highlights the prediction interface for entering new patient data and receiving real-time risk assessments.

Annexure G: System Architecture and Deployment

Describes the overall system architecture including client devices, central server, and communication protocols.

- Explains deployment options such as local hospital servers or cloud-based platforms.

- Discusses security measures implemented during communication between clients and server.

- Details scalability considerations for increasing number of clients or data volume.

Annexure H: Ethical and Regulatory Considerations
Summarizes relevant healthcare data privacy regulations such as GDPR and HIPAA.

- Discusses how federated learning and differential privacy help comply with these regulations.

- Highlights ethical concerns around data use, patient consent, and transparency.

- Suggests best practices for responsible AI deployment in healthcare environments.