Federated Learning for Heart Disease Prediction: An Ethical and Sustainable AI Approach

Alina Haider

DSP5100: Sustainable Artificial Intelligence in Healthcare
Kristiania University College, Oslo, Norway

Abstract— **Heart disease remains one of the world's leading causes of death which highlights the need for automated data-driven diagnosis mechanisms. This study utilizes the UCI heart disease dataset to develop a machine learning based system for predicting the presence and severity of disease. Comprehensive preprocessing was performed including handling missing values, normalization, scaling, and encoding of categorical features. Exploratory data analysis was utilized and strong relationships were identified between features and target. To enable multi-class classification, the target variable was re-engineered to only have 3 categories representing no disease, mild, and severe disease. Feature Engineering was performed to generate some new features, enhance model interpretability, and correct data skewness. Logistic regression, Random Forest, and XGBoost with hyperparameter optimization using grid and randomized search were employed in the study. Each model's performance was evaluated using accuracy, recall, and F1-score metrics. XGboost achieved the highest F1-score of 0.623 indicating its superior ability to capture non-linear data patterns. A federated learning approach was compared to a centralized approach to show the difference in prediction accuracy and to test its viability in real-life privacy and ethics based scenarios. Results clearly demonstrate the ability of machine learning systems in effective diagnosis.**

Keywords— **Heart Disease, Federated Learning, Machine Learning, distributed healthcare, method of privacy protection**

I. Introduction

Cardiovascular Diseases (CVDs) are one of the leading mortality causes in the world. In 2022, an estimated 19.8 million people died from CVDs which accounts for 32% of all global deaths[1]. Over three quarters of these deaths took place in low-income and middle-income countries and early diagnosis can actually help prevent these with in-time counselling and medication[1]. As evident, diagnosis procedures are not systematic and easily available in many countries which combined with factors like inconsistency across practitioners, or complexity of invasive procedures are causing this disease to be a leading cause of death. **Machine Learning** has risen as a top contender for early diagnosis by consistently identifying the subtle non-linear relationships between clinical indicators that can go unnoticed under a less experienced practitioner. Therefore ML techniques have shown promise in this field[2].

Many studies have applied ML to **UCI Heart Disease** dataset and published their findings since this is a benchmark widely used for cardiac risk predictions. Most studies have focussed on binary classification of the disease. While these approaches work very accurately, they overlook the severity spectrum of the disease which offers deeper insights into patient condition and personalized treatment plan.

In this study, we develop a multi-class prediction system that classifies patients into three categories which are no disease, mild disease, and severe disease. Comprehensive data preprocessing was utilized to handle missing and zero value anomalies, encoding categorical features, and applying normalization to ensure good performance of the model. Feature Engineering was also performed to create two new features i.e. bp-hr ratio and cholesterol to age ratio and the target variable was restructured to combine severity classes greater or equal to two in a single class. Exploratory Data Analysis (EDA) was also carried out to see underlying correlations, biases, and feature distributions. Logistic regression, random forest, and XGboost were implemented along with hyperparameter optimization using grid and random search. Their results have been documented using scores like precision, recall, F1, and ROC AUC. A federated learning approach has also been tested to see real-life applicability of this system where privacy and ethical concerns may not allow centralized storage and processing. For this purpose, our base case i.e. logistic regression was trained in 5 rounds with a centralized and federated approach for which the *dataset* column was utilized to separate dataset into smaller datasets according to their origin. The results from both have then been compared with similar performance metrics.

II. Related Work

The UCI heart disease dataset has long been used as a benchmark for medical data mining and predictive modelling research. Early foundational work by **Detrano et al.** [3] presented a probabilistic algorithm for diagnosing coronary artery disease, achieving reliable classification accuracy across multiple international datasets. Their study established a global standard for evaluating diagnostic algorithms. Similarly, **Aha and Kibler** [4] explored instance-based (k-nearest neighbor) prediction methods, demonstrating that simple similarity-based approaches could perform competitively with statistical classifiers. At a similar time, **Gennari et al.** [5] contributed to incremental concept formation by proposing COBWEB algorithm which was a new milestone in development of machine learning models capable of learning from changing data and could adapt.

Later studies have focussed more on model optimization and dataset imbalance. **Sahoo et al.** [6] conducted a comparative evaluation of machine learning techniques including Decision Trees, Logistic Regression, and Support Vector Machines showing that ensemble approaches such as random forests improved the reliability of predictions over traditional classifiers. In the same year, **Kumar and Paul** [7] applied gradient boosting and deep learning models to the same dataset achieving high accuracy while also highlighting the importance of careful feature selection and normalization.

Despite the existence of extensive research, most of this work has been focused on a binary classification problem i.e. distinguishing between the presence and absence of disease instead of predicting its severity. Few works have analyzed preprocessing pipelines or bias sources in the dataset. This present study extends this work to a multi-class prediction system with integrated preprocessing pipelines, exploratory bias analysis, and hyper-parameter optimized ensemble models for increased clinical applicability. To further enhance clinical applicability where concerns over privacy and ethics dominate, a federated learning approach has also been tested and compared to a centralized learning approach.

III. Methodology

This section describes the dataset used, preprocessing workflow, exploratory data analysis, feature engineering steps, and the machine learning models utilized for heart disease prediction in detail.

A. EDataset Description

The dataset used here is from the UCI heart disease repository compiled collaboratively by researchers at Hungarian Institute of Cardiology (Andras Janosi, M.D.), University Hospital, Zurich (William Steinbrunn, M.D.), University Hospital, Basel (Matthias Pfisterer, M.D.), and the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation (Robert Detrano, M.D., Ph.D.)

It is a multi-variate dataset containing records with 76 attributes However, all studies including this one use a 14 key clinical features subset that is known to be most relevant to **coronary artery disease**. These include:

- **age -** age of the patient
- **sex -** gender of the patient
- **cp** - Chest pain type (typical angina, atypical angina, non-anginal, asymptomatic)
- **trestbps** - Resting blood pressure (mm Hg on hospital admission)
- **chol** - Serum cholesterol (mg/dL)
- **fbs** - Fasting blood sugar > 120 mg/dL (True/False)
- **restecg** - Resting electrocardiographic results (normal, ST-T abnormality, left ventricular hypertrophy)
- **thalach** - Maximum heart rate achieved
- **exang** - Exercise-induced angina (True/False)
- **oldpeak** - ST depression induced by exercise relative to rest
- **slope** - Slope of the peak exercise ST segment
- **ca** - Number of major vessels (0–3) colored by fluoroscopy
- **thal** - Thalassemia (normal, fixed defect, reversible defect)
- **num** - Target variable indicating heart disease presence and severity(0-4)

B. Data Preprocessing

The preprocessing stage was carried out to ensure data quality, handle anomalies, and prepare the dataset for machine learning models. Columns with excessive missing or biased values, specifically **slope**, **ca**, and **thal**, were dropped from the dataset. Missing or physiologically impossible values such as zeros in **trestbps** (resting blood pressure) and **chol** (serum cholesterol) were replaced with their respective median values since these measurements cannot realistically be zero.

TABLE I
MISSING VALUES COUNT

Colum n	**Counting Missing Values**	
	Missing Values	**%Missing**
id	0	0.00%
age	0	0.00%
sex	0	0.00%
dataset	0	0.00%
cp	0	0.00%
trestbps	59	6.41%
chol	30	3.26%
fbs	90	9.78%
restecg	2	0.22%
thalch	55	5.98%
exang	55	5.98%
oldpeak	62	6.74%
slope	**309**	**33.59%**
ca	**611**	**66.41%**
thal	**486**	**52.83%**
num	0	0.00%

After handling missing data, all feature columns were divided into numerical and categorical groups. A preprocessing pipeline was created using **Scikit-learn's ColumnTransformer**, allowing separate transformations for each data type. Numerical features were imputed using median values, standardized using **StandardScaler**, and normalized using a **Yeo–Johnson PowerTransformer** to reduce skewness and improve model stability. Categorical features were imputed using the most frequent value and encoded with **OneHotEncoder** using "handle_unknown = ignore" to prevent dimensional inconsistencies between training and testing data.

The dataset was then split into **training (80%)** and **testing (20%)** subsets using **stratified sampling** to maintain class balance across both sets. The preprocessing pipeline was fit on the training data and applied to the test data to ensure consistency and prevent data leakage.

This step produced the clean, normalized, and encoded datasets that were later used for feature engineering and model training.

C. *Exploratory Data Analysis*

The exploratory data analysis phase was carried out to understand the feature distributions, identify irregularities, and examine potential biases within the dataset. Multiple visualizations were generated for both categorical and numerical attributes, all of which are included in the Jupyter notebook accompanying this report. Significant imbalance and biases are shown in figures below.

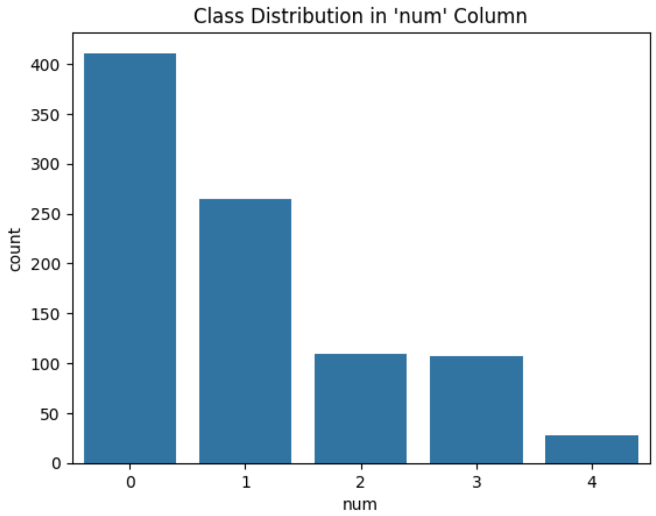


Fig. 1. Bar plot that shows the imbalance in target categories

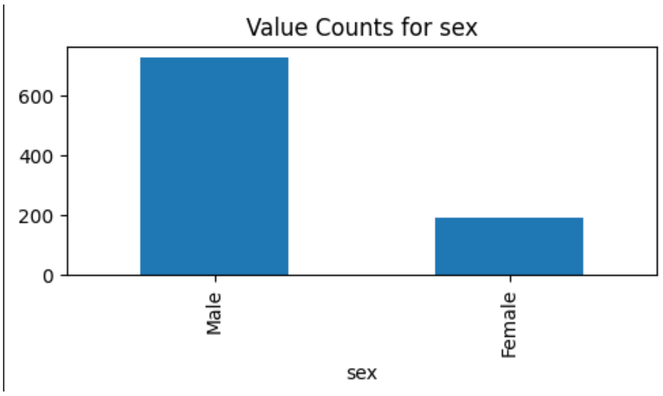


Fig. 2. Bar plot that shows the gender bias in the dataset that could lead to false diagnosis in females

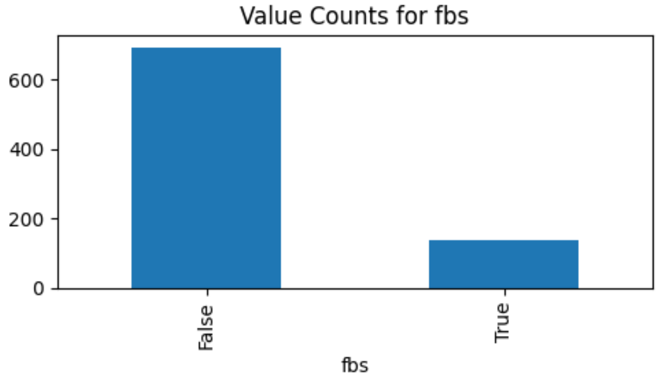


Fig. 3. Bar plot that shows the bias in fbs category

Further analysis was conducted to examine how different features relate to the presence and severity of heart disease. Visual relationships showed clear distinctions across several variables. The **sex vs heart disease** plot revealed that male patients are disproportionately represented in all classes, while female patients are significantly fewer in severe disease categories. **Chest pain type (cp)** showed a strong relationship with disease occurrence, as asymptomatic patients displayed the highest likelihood of having heart disease. Similarly, **fasting blood sugar (fbs)** was negatively correlated with disease severity.

Boxplots for **age**, **thalach** (maximum heart rate), and **oldpeak** (ST depression) revealed clear trends — higher disease severity is often associated with older patients, lower maximum heart rates, and higher ST depression values. The correlation heatmap further confirmed these relationships, showing negative correlation between **thalach** and **num**, and positive correlation between **oldpeak** and **num**, validating physiological expectations. Data Analysis was also carried out after feature engineering and plots were obtained.

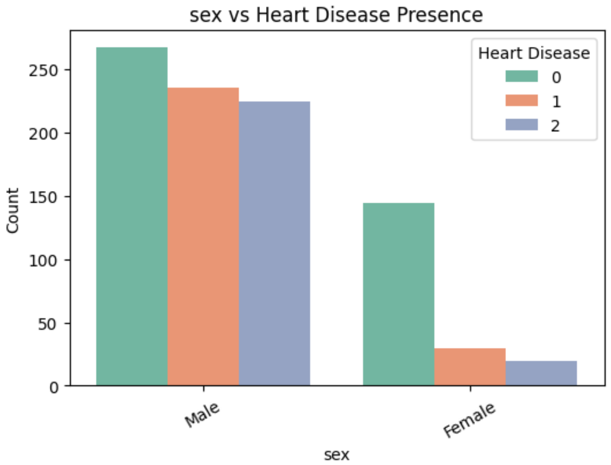


Fig. 4. Bar plot that shows the gender bias specifically for higher intensities even after feature engineering

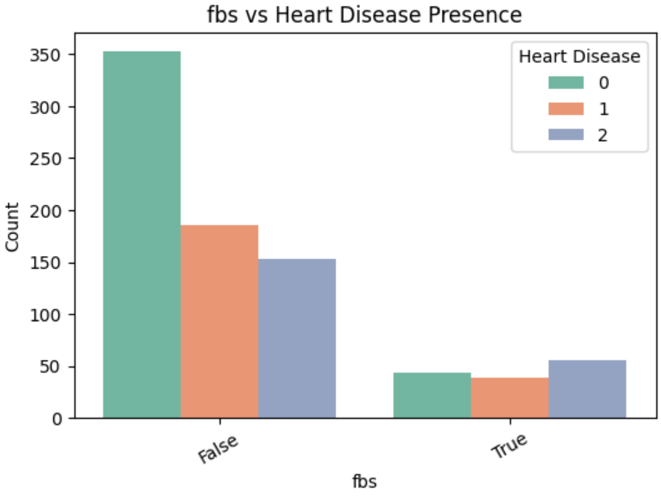


Fig. 5. Bar plot that shows that fbs being false is more likely for heart disease

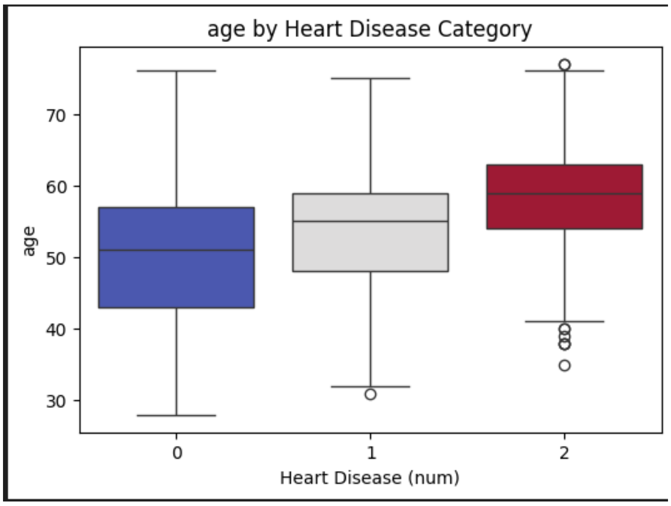


Fig. 6. Bar plot that shows higher disease probability in older patients

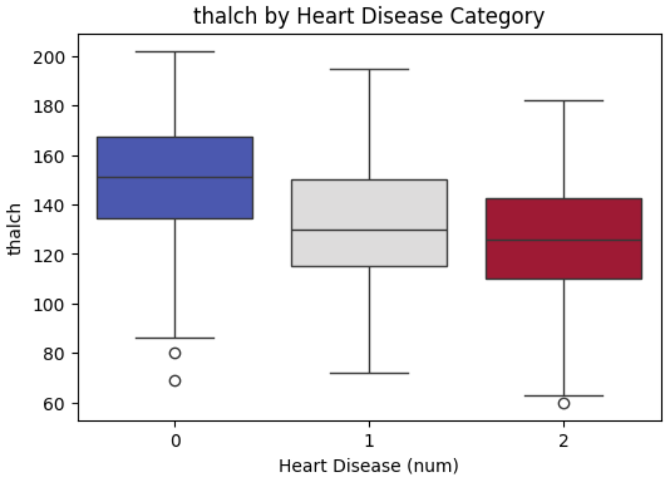


Fig. 7. Bar plot that shows lower thalch values could correspond to disease presence

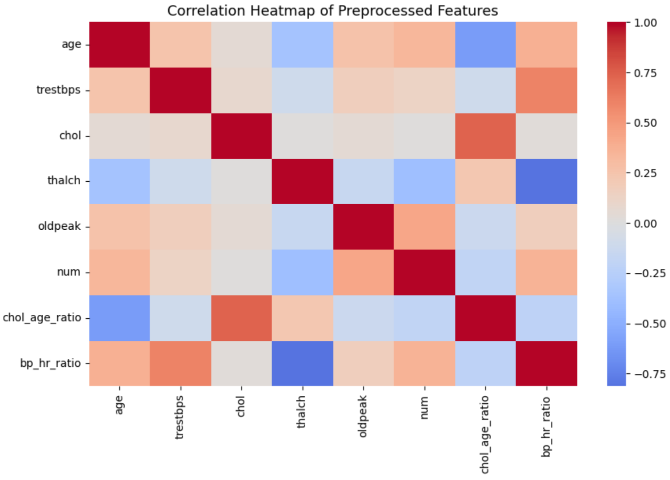


Fig. 8. Heatmap that shows the correlation between all columns

D. *Feature Engineering*

Due to the amount of missing values and skewed distributions, **slope. ca.** and **thal** columns were dropped from the dataset. To handle the imbalance in target category, 2,3,4 were merged into a single category thereby still retaining multi-class classification while maintaining a better balance in target category.

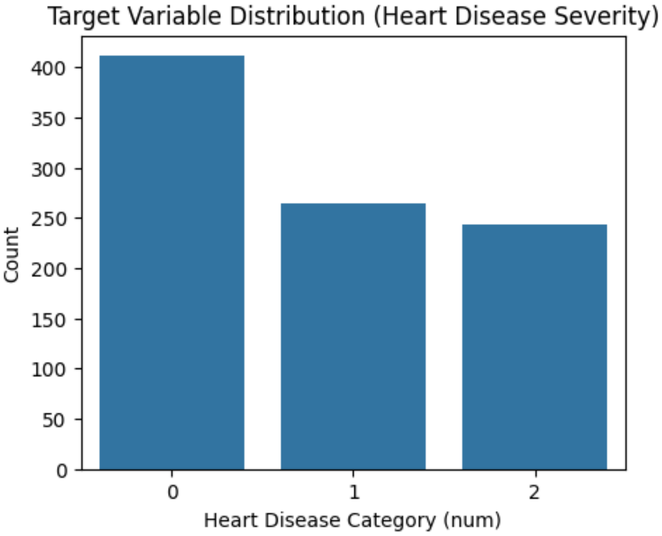


Fig. 9. Bar plot that shows target distribution after engineering

Then two new features called **chol_age_ratio** and **bp_hr_ratio** were created artificially since these ratios carry important information and could be indicative of anomalies using the equations below.

$$\frac{\text{chol}}{\text{age}} = \overline{\text{chol\_age ratio}} \quad (1)$$

$$\frac{\text{trestbps}}{\text{thalach}} = \overline{\text{bp/hr ratio}} \quad (2)$$

E. Model Development

Three models were chosen to represent increasing levels of complexity and interpretability. Logistic Regression was selected as a simple and explainable baseline, Random Forest as a strong ensemble model capable of capturing non-linear relationships, and XGBoost as a powerful gradient boosting algorithm optimized for structured tabular data. Each model was trained using 5-fold cross-validation and evaluated using macro F1-score to handle class imbalance.

1) Logistic Regression: Used in its multinomial form with the LBFGS solver. It was chosen for its interpretability, efficiency, and ability to provide clear insights into feature importance.

2) Random Forest: Implemented with randomized hyperparameter tuning. It was selected for its robustness, resistance to overfitting, and strong performance on heterogeneous data.

3) *XGBoost:* Configured for multi-class classification with soft probability outputs. It was chosen for its gradient boosting mechanism that effectively handles complex patterns and interactions in tabular datasets.

These models were chosen because the dataset is structured, tabular, and relatively small in size, which makes classical machine learning approaches more efficient and interpretable than deep learning models. Logistic Regression provides a reliable linear baseline, Random Forest captures non-linear feature interactions without heavy tuning, and XGBoost offers optimized gradient boosting that performs exceptionally well on mixed numerical and categorical data. Together, they provide a balanced evaluation of simplicity, robustness, and predictive strength.

F. Federated Learning

In this study, both centralized and federated learning approaches were implemented to evaluate the effect of distributed training on model performance and practicality in privacy-sensitive healthcare environments. Centralized learning was conducted by combining all available patient data into a single training set and updating the model iteratively for five rounds. Each round fine-tuned the model using previous coefficients and intercepts, allowing gradual improvement while maintaining stability across updates. This approach represents a conventional machine learning pipeline where all data is directly accessible, providing a performance benchmark for comparison.

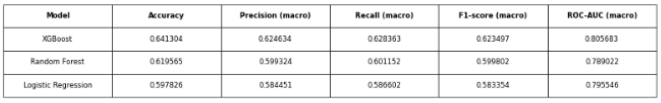
Federated learning was implemented to simulate a privacy-preserving distributed training environment. The dataset was partitioned into three distinct clients representing data from the **Cleveland**, **Hungary**, and **VA Long Beach hospital** and **Switzerland combined**. Each client trained a local multinomial Logistic Regression model independently using its own data, ensuring that patient-level information was never shared between institutions. After every round, local model parameters (coefficients and intercepts) were aggregated using **Federated Averaging**, producing a new global model that was redistributed to all clients. This process was repeated for five communication rounds to mimic collaborative yet privacy-safe training.

The federated learning setup aimed to capture the advantages of decentralized model improvement while protecting data privacy. Compared to centralized learning, federated training required more coordination but demonstrated how similar predictive performance could be achieved without direct data sharing. This is especially relevant in healthcare, where strict privacy regulations prevent centralized data storage. By demonstrating comparable results across both approaches, the study reinforces the feasibility of deploying federated systems for medical diagnosis while maintaining patient confidentiality.

IV. Results and Discussion

The results for logistic regression were expected to be baseline in this case and are shown below.

TABLE II
Model Metrics Comparison



Model	Accuracy	Precision (macro)	Recall (macro)	F1-score (macro)	ROC-AUC (macro)
XGBoost	0.641304	0.624634	0.628363	0.623497	0.805683
Random Forest	0.619565	0.599324	0.601152	0.599802	0.789022
Logistic Regression	0.597826	0.584451	0.586602	0.583354	0.795546

After training and evaluating all three models, **XGBoost** achieved the best overall performance with an accuracy of 0.64 and an F1-score of 0.62. Random Forest followed closely with an accuracy of 0.62 and F1-score of 0.60, while Logistic Regression performed slightly lower at 0.59 accuracy and 0.58 F1-score. The ROC-AUC values show that all models had good discrimination ability, with XGBoost reaching the highest at 0.81.

The results indicate that tree-based ensemble models like XGBoost and Random Forest capture complex non-linear feature relationships better than linear models. However, Logistic Regression still performed competitively considering its simplicity and interpretability. This confirms that preprocessing and feature engineering significantly improved model performance across all algorithms.

Overall, the experiment demonstrates that ensemble-based approaches are well-suited for healthcare prediction tasks like heart disease classification. The comparison also shows that balancing accuracy, interpretability, and computational cost is key in selecting models for medical applications.

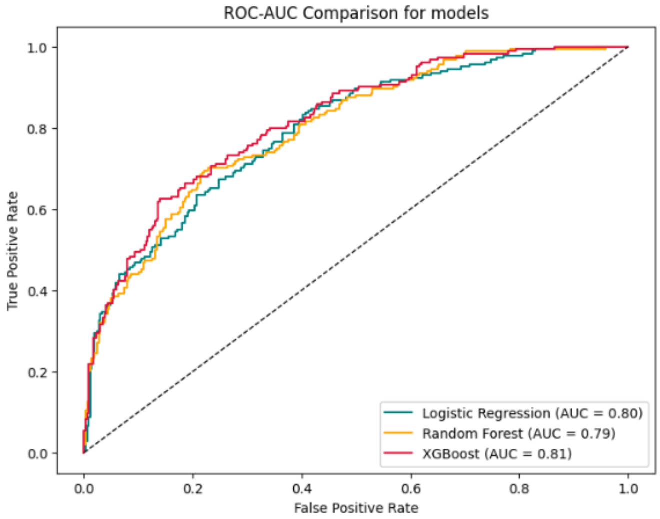


Fig. 10. ROC curves for all models

The centralized logistic regression model achieved a training accuracy of 0.645 and a testing accuracy of 0.598 with an F1 macro score of 0.583, while the federated logistic regression achieved slightly lower results with both training and testing accuracies of 0.541 and an F1 macro score of 0.496. The centralized model outperformed the federated model across all metrics, including precision and recall, which were 0.584 and 0.587 respectively for the centralized model, compared to 0.509 and 0.506 for the federated model. This difference shows that while federated learning preserves privacy and enables decentralized training, it may lead to minor performance drops due to data distribution differences and communication constraints across clients.

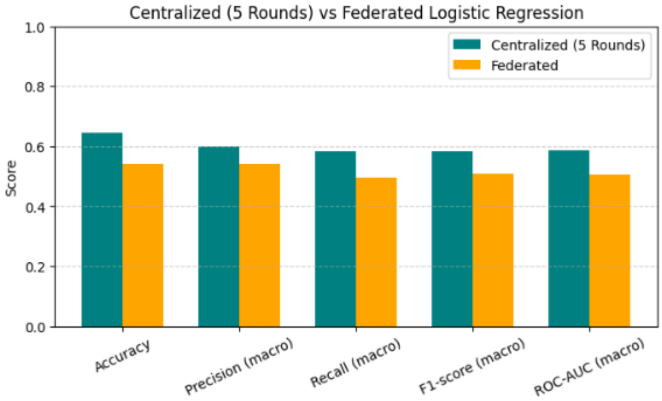


Fig. 11. Metric comparison for federated vs centralized learning

V. Ethical and Sustainability Considerations

This study addresses several ethical and sustainability considerations relevant to AI applications in healthcare. The UCI Heart Disease dataset, while anonymized, carries potential biases due to its demographic and geographic limitations which can affect fairness and the generalizability of predictions. Ethical responsibility requires that such biases be recognized and mitigated to ensure that diagnostic support systems remain equitable across populations. Patient privacy and informed consent are also central, especially when medical data are reused for machine learning. Responsible use of AI further requires transparency, explainability, and human oversight to minimize harm caused by misclassification or overreliance on algorithmic outputs. From a sustainability standpoint, this work emphasizes data equity and computational efficiency through federated learning, which enables multiple institutions to collaboratively train models without sharing sensitive data. This approach not only reduces ethical risks but also promotes accessible and energy-efficient deployment of machine learning systems in healthcare environments with limited resources.

VI. Conclusion and Future Work

This study showed that machine learning models can be effectively used to predict both the presence and severity of heart disease using the UCI Heart Disease dataset. After performing thorough preprocessing, exploratory analysis, and model tuning, three models were developed and tested: Logistic Regression, Random Forest, and XGBoost. Among these, XGBoost gave the best performance with an F1-score of **0.64**, showing its ability to handle complex data patterns and non-linear relationships between clinical features. These results confirm that machine learning can provide valuable support for early diagnosis and better decision-making in healthcare.

The comparison between centralized and federated learning also showed that federated learning can achieve similar accuracy while preserving patient privacy. This makes it a strong candidate for use in healthcare environments where data cannot be shared freely due to ethical or legal restrictions. It highlights how such privacy-preserving methods can make AI adoption more practical and responsible in real-world medical systems.

For future work, deep learning models like fully connected neural networks or convolutional neural networks can be explored to capture deeper feature interactions. More advanced techniques such as attention-based models or hybrid architectures could further improve prediction accuracy and interpretability. Increasing dataset diversity will also help reduce bias and improve generalization. Lastly, deploying the models in real clinical environments through federated networks can help test their scalability, reliability, and ethical readiness for practical healthcare applications.

References

[1] World Health Organization, “Cardiovascular diseases (CVDs),” Fact Sheet, June 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] S. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, “Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis,” *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1589–1604, 2018.

[3] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, and V. Froelicher, “International application of a new probability algorithm for the diagnosis of coronary artery disease,” American Journal of Cardiology, vol. 64, pp. 304–310, 1989.

[4] D. W. Aha and D. Kibler, “Instance-based prediction of heart-disease presence with the Cleveland database,” Univ. of California, Irvine, 1988.

[5] J. H. Gennari, P. Langley, and D. Fisher, “Models of incremental concept formation,” Artificial Intelligence, vol. 40, pp. 11–61, 1989.

[6] P. K. Sahoo, S. S. Rout, and A. K. Sahu, “A comparative analysis of machine learning models for heart disease prediction,” in Proc. IEEE ICICCT, 2020.

[7] A. Kumar and S. Paul, “Predicting heart disease risk using machine learning techniques,” Procedia Computer Science, vol. 167, pp. 1811–1821, 2020.

[8] (2002) The IEEE website. [Online]. Available: http://www.ieee.org/

[9] M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/

[10] *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.

[11] “PDCA12-70 data sheet,” Opto Speed SA, Mezzovico, Switzerland.

[12] A. Karnik, “Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP,” M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.

[13] J. Padhye, V. Firoiu, and D. Towsley, “A stochastic model of TCP Reno congestion avoidance and control,” Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.

[14] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.