Centralized stacking ensemble and Federated Learning Models for Heart Disease Prediction with SHAP

S Rama Shesha Sai[1]
ramasheshasai_s@srmap.edu.in

Rohan Tunikipati[2]
rohan_tunikipati@srmap.edu.in

David Raju Boddupalli[3]
davidraju_boddupalli@srmap.edu.in

Srinivas Jilla[4]
Srinivas_jilla@srmap.edu.in

Lipismita Panigrahi 0000−0003−2952−46285
lipismita.p@srmap.edu.in

[1-5]Department of Computer Science and Engineering, SRM UNIVERSITY, Amaravati, Andhra Pradesh, India

***Abstract*—Cardiovascular diseases (CVDs) remain a leading global health risk, necessitating accurate and efficient diagnostic techniques. This study proposes a hybrid deep learning-based architecture for heart disease prediction using the publically accessible Heart Disease dataset, which includes 920 patient records and 13 significant clinical characteristics. The suggested model, known as Power Boost Ensemble, uses a stacking technique to merge four distinct base learners: Random Forest, Extra Trees, Gradient Boosting, and Logistic Regression. A Ridge Classifier serves as the meta learner in this configuration, gathering predictions from each base learner. With a test accuracy of 85% using 10-fold cross validation, the stacked ensemble exhibits good generalization and consistent performance across all significant assessment criteria. Shapley Additive Explanations (SHAP) are used to understand how the meta model develops its predictions in order to improve interpretability and clarity. The SHAP results show that the model's conclusions are significantly influenced by important clinical parameters including ca (number of main vessels), cp (kind of chest pain), thal (thalassemia), and oldpeak (ST depression). All things considered, the Power Boost Ensemble offers a dependable, comprehensible, and reproducible approach to cardiac sickness prediction, making it appropriate for upcoming clinical applications based on artificial intelligence.**

***Index Terms*—Cardiovascular Disease, Heart Disease Prediction, Ensemble Deep Learning, Explainable AI, Interpretability, Multi-Layer Perceptron, Federated Learning Models.**

I. Introduction

A. Motivation

Cardiovascular diseases continue to be the world's leading cause of death, there is a huge need for diagnostic models that are accurate, understandable, and sensitive of patient privacy [1]. According to recent studies [2], deep learning (DL) and machine learning (ML) are becoming more and more useful in identifying early indications of cardiovascular risk. Techniques that combine interpretability approaches like SHAP with ensemble learning have also improved transparency and prediction consistency in clinical settings [3]. Healthcare platforms have further enhanced decision support by incorporating explainable learning pipelines into clinical processes [4], [5]. Federated learning systems, such as FedeHR, meet strict privacy requirements by showing how various institutions may train shared models without sharing raw patient data [1]. However, despite these developments, studies employing the Heart Disease dataset continue to reveal noticeable differences in model accuracy, raising concerns about stability and reliability in real-world medical applications [6]. This study proposes a stable predictive framework that addresses these problems by combining a **centralized stacking ensemble** utilizing a Ridge Classifier as the meta model and a **federated learning** configuration that uses a Multi-Layer Perceptron (MLP) trained using FedAvg. Further, by using Shapley Additive Explanations (SHAP) [7], the integrated approach aims to increase prediction accuracy, protect data confidentially across institutions, and give clearer interpretability. The main objective is to establish a standard for cardiac disease prediction that is safe, clear, and clinically reliable.

B. Origin of the Problem

- When used in real clinical settings, current cardiac disease prediction models often rely on tiny datasets, which leads to inconsistent performance and reduced reliability [4].
- Additionally, doctors find it challenging to fully trust or confirm predictions provided by these systems since most machine learning models are not sufficiently interpretable [5].
- Because they offer effective learning while upholding strict data privacy, federated and ensemble-

based techniques have become increasingly important in decentralized data contexts [1].
- The need for safer and more transparent AI-driven healthcare solutions is highlighted by the ongoing gaps in clinical decision support caused by the lack of a consistent and easily comprehensible prediction framework [2].

C. Contribution and Outline of Paper

- Proposed a **Power Boost Ensemble**, a centralized stacking model that combines Random Forest, Extra Trees, Gradient Boosting, and Logistic Regression, with a Ridge Classifier acting as the meta-learner for predicting heart disease with a test accuracy of **85%** through 10-fold cross-validation, demonstrating reliable generalization and stable results across various evaluation measures described in II-B.
- Further Enhanced the federated Model with the MLP Achieving a test accuracy of 80%.
- Conducted an comparison between the centralized and federated models analyzing accuracy, precision, recall, F1-score, and loss patterns and further compared the proposed model with state-of-the-art existing models.
- Delivered a reproducible workflow that effectively unifies **high predictive performance, privacy preservation, and interpretability**, demonstrating its suitability for real-world medical applications.

The outline of the paper is organized as follows. Section II explains the materials and methods used in this study and provides a detailed description of the complete architecture for both the centralized stacking ensemble and the federated learning framework, along with their corresponding algorithms. Section III provides the experimental setup, evaluation metrics, and comparative performance analysis across centralized and federated models. Finally Section IV conclude the paper.

II. Materials and Methods

A. Demographics and Preprocessing of Data

The Heart Disease dataset(UCI Heart Disease repository), a widely used benchmark for cardiovascular prediction, is employed in this study [8], [9]. It includes 920 records from four institutions: VA Long Beach, the Cleveland Clinic, the Hungarian Institute of Cardiology, and hospitals in Zurich and Basel. Initially, preproced the datasets like :the missing values marked with "?" were handled using median imputation, and numerical features were standardized using StandardScaler. In the federated setup, each site kept its data local and shared only model parameters, maintaining privacy throughout training. Although the full dataset contains 76 attributes, this work focuses on 14 key clinical variables that show strong relevance to UCI heart disease outcomes shown in Fig. 1.

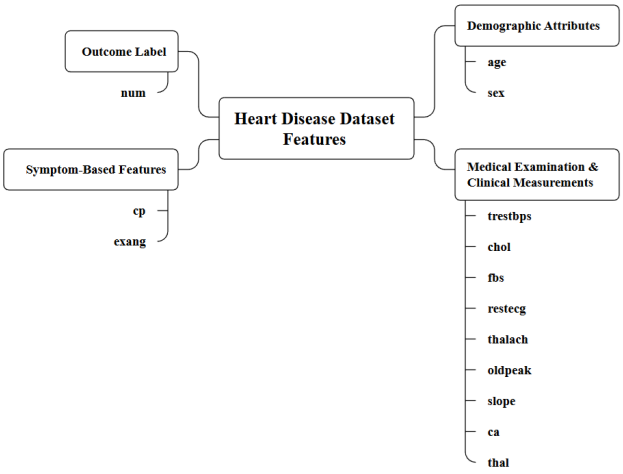

Fig. 1. Heart Disease Features

$$D = D_1 \cup D_2 \cup \cdots \cup D_K, \quad D_i \cap D_j = \emptyset \qquad (1)$$

As shown in (1), this represents the K-fold partitioning of the dataset, where the entire dataset D is divided into K mutually exclusive subsets. Each subset serves once as a validation fold, while the remaining folds are used for training, ensuring unbiased and robust model evaluation.

B. Proposed Model

This proposed model is of two phases and in the first phase we propsoed an centralied model by combine RF, ET ,GB, LR and in the Second Phase we enhanced the Federated model. The description of both model is given below (all the codes can be found in https://github.com/ramasheshasai/UROP$_{P}ROJECT$.

*1) **Centralized Stacking Ensemble Framework**:* To improve the reliability and interpretability of heart disease classification, a **stacking ensemble model** is designed for the centralized learning environment. Ensemble techniques combine the strengths of diverse algorithms tree-based models capture complex nonlinear interactions, while linear models improve stability and offer clearer interpretability [2], [6].

The proposed framework integrates four tuned base classifiers Such as RF, ET, GB, and LR trained on the pre-processed Heart Disease dataset [1]. The models were configured as follows:

1) **Random forest (RT):** Extra Trees constructs an ensemble of decision trees, but introduces Some randomness to further reduce variance. Instead of selecting the best split threshold, RT randomly samples split points and evaluates them, which creates more diverse and less correlated trees.

Random forest chooses the best split based upon the impurity measures from a random subset of features. The final prediction is obtained by majority voting across all trees as given in (2).

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_K(x)\} \quad (2)$$

2) **Extra Trees (ET):** Extra Trees further increase randomness by selecting split thresholds uniformly at random rather than computing the optimal split. This creates highly diverse trees and speeds up training. A node split follows: $s =$ random threshold in feature f
3) **Gradient Boosting (GB):** Gradient Boosting builds the model in stages, where each new weak learner attempts to correct the residual errors of the previous learners. This iterative refinement improves accuracy. The update step is: $F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x)$
4) **Logistic Regression (LR):** Logistic Regression estimates the probability of a positive class by applying a sigmoid function to a linear combination of input features. The prediction model is expressed in (3).

$$P(y = 1 \mid x) = \sigma(w^\top x) = \frac{1}{1 + e^{-w^\top x}} \quad (3)$$

In order to create a cohesive predictive system, the centralized learning architecture uses a hybrid stacking ensemble, as shown in Fig. 2.We choose the name as power boost ensemble because combining learners allows to capture both linear and nonliner clinical patterns. Every model offers a different approach to learning: logistic regression offers stability and a clear decision boundary, whereas tree-based algorithms capture intricate nonlinear patterns. Each base learner is trained using 10-fold cross-validation to efficiently aggregate these strengths, resulting in out-of-fold predictions that together create a meta-feature matrix. This matrix gives the meta-learner a better and more insightful depiction of how each model behaves across unseen data. The stacked ensemble can then combine both linear and nonlinear interactions by using a Ridge Classifier as the meta-model, as seen in Step 5 of the Algorithm 1. A weighted linear aggregation of the base learner outputs found in 4 produces the final prediction.

$$\hat{y} = \text{sign}\left(\sum_{i=1}^{M} w_i \, h_i(x)\right) \quad (4)$$

where $h_i(x)$ is the prediction from the i-th model, w_i is the coefficient learned through Ridge regularization, and M is the total number of base models.

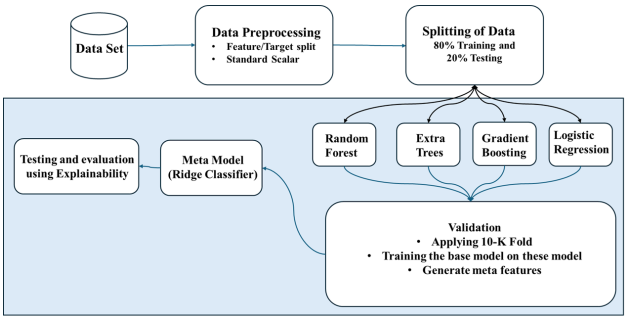

Fig. 2. Proposed Centralised Framework.

*2) **Enhanced Federated Learning Architecture Using MLP**:* The framework of this enhanced model is shown in Fig. 3. The FL setup maintains all patient data locally while enabling collaborative cardiac disease prediction across the Cleveland, Hungarian, Switzerland, and VA Long Beach datasets in compliance with HIPAA and GDPR rules [10]. Each client uses median imputation, normalization, and an 80–20 stratified split prior to training an identical MLP with three fully connected layers (64 ReLU, 32 ReLU, and 2 Softmax), dropout of 0.2, and the Adam optimizer with a learning rate of 10^{-3}. Each round of local training consists of three epochs and a batch size of sixteen.

$$w^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{N} w_k^{(t)}, \quad (5)$$

Clients only share model parameters after each round, and FedAvg [11] is employed to maintain global coordination. The server uses (5), weighting updates by the number of local samples, to ensure balanced contributions. The modified global model is then sent to every client, and this procedure is repeated during communication rounds. The final

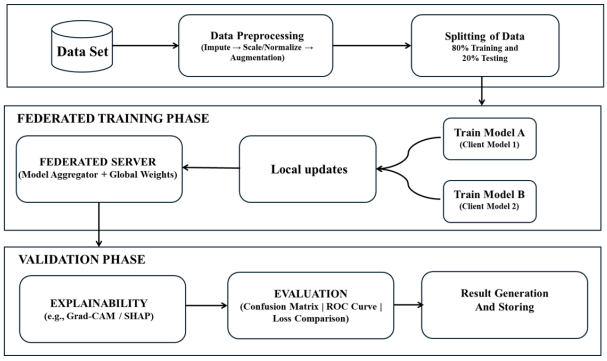

Fig. 3. Enhanced Federated Framework.

evaluation uses a composite test set created from the held-out data from all four colleges

Algorithm 1 Proposed Centralized Stacking Ensemble Power Boost model

Input: Dataset D, base models M (RF, ET, GB, LR), and the number of folds K.
Output: Final trained ensemble model, predicted outputs, accuracy scores, and SHAP explanations. explanations.
(1) $P \leftarrow$ **Preprocessing**(D)
(2) $S \leftarrow$ **SplitDataset**(P)
(3) $F \leftarrow$ **KFoldPartition**(S, K)
(4) $Z \leftarrow$ **GenerateMetaFeatures**(M, F)
(5) $R \leftarrow$ **TrainRidgeClassifier**(Z)
(6) $\hat{Y} \leftarrow$ **Predict**(R, test_samples)
(7) $A \leftarrow$ **EvaluateAccuracy**($\hat{Y}$, groundtruth)
(8) $C \leftarrow$ **ComputeConfusionMatrix**($\hat{Y}$, groundtruth)
(9) $E \leftarrow$ **SHAPExplain**(R)
(10) **Return:** ensemble model, predictions, accuracy, and SHAP explanations

Each institution separately preprocesses its local dataset by normalizing features and imputing missing values in **Step 1** of Algorithm 2. The global model parameters, which are the starting point for all clients, are initialized in **Step 2**. The collaborative training process is structured by a series of communication rounds that the system enters in **Step 3**. While maintaining the privacy of all raw records, **Step 4** permits each client to do local MLP training using its own data throughout each round. After local updates are finished, **Step 6** uses FedAvg to aggregate the model parameters so that all institutions' contributions are reflected in the global model.In order to evaluate prediction performance and overall model behavior, the final trained global model is then assessed in **Steps 8** and **9** on a shared test set. The algorithm is represented by Fig.reffflow. There are two main advantages to this federated system. By maintaining all data inside each hospital, it safeguards patient privacy while enabling the global model to learn from a variety of clinical patterns, lowering site bias and enhancing generalization. Because of these advantages, federated learning is a good fit for accurate cross-institutional medical prediction [12], [13].

III. Experimental Evaluation and Discussion

A machine with an Intel Core i7 CPU, 16GB RAM, an NVIDIA RTX 3060 GPU, and Python 3.10 was used for all of the tests. `scikit-learn` was used to construct centralized models, and `PyTorch` was used to develop the federated framework. With a batch size of 16 and a learning rate of 10^{-3}, each model was trained for 50 epochs. Accuracy, precision, recall, and F1-score,all commonly used classification metrics for clinical and healthcare prediction tasks, were used to assess the model's performance [14]–[17].

Algorithm 2 Federated Learning Algorithm (MLP with FedAvg)

Input: Client datasets $D = \{D_1, D_2, D_3, D_4\}$, initial global parameters θ_0, communication rounds R.
Output: Final global model and evaluation measures.
Algorithm:
(1) $P \leftarrow$ Preprocessing(D) // impute missing values and normalize per client
(2) $\theta \leftarrow$ InitializeGlobalParameters(θ_0)
(3) **for** each round $r = 1$ to R **do**
(4) **for** each client k **do** $\theta_k \leftarrow$ TrainMLP(D_k, θ) // local training
(5) **end for**
(6) $\theta \leftarrow$ FedAvg($\{\theta_k\}$) // aggregate by sample size
(7) **end for**
(8) $Y \leftarrow$ Predict(θ, global test set)
(9) $A \leftarrow$ EvaluateAccuracy(Y, groundtruth)
(10) **Return:** Final model prediction and accuracy.

A. Quantitative Results

*1) **Centralized Stacking Ensemble Performance**:* The complimentary strengths of the four basic models—RF, ET, GB, and LR—trained with 10-fold cross validation are reflected in the performance in Table I. LR offers a stable linear border, GB enhances challenging instances through its boosting mechanism, and RF and ET introduce rich nonlinear patterns. By utilizing L2 regularization to combine these outputs, the Ridge meta learner avoids over-contribution from any one model. An accuracy of 0.85 with closely matched precision(0.85) and recall(0.85) across both classes is the outcome of this balanced integration.

*2) **Interpretation of Centralized Confusion Matrix** :* The confusion matrix in Fig. 4 shows 30 true negatives, 21 true positives, and a small number of

TABLE I
Performance Metrics for Centralized Stack Ensemble

Metric	Class 0: No Illness	Class 1 : Illness	Overall
Precision	0.88	0.81	0.85
Recall	0.86	0.84	0.85
F1-Score	0.87	0.82	0.85
Accuracy	**0.85**		

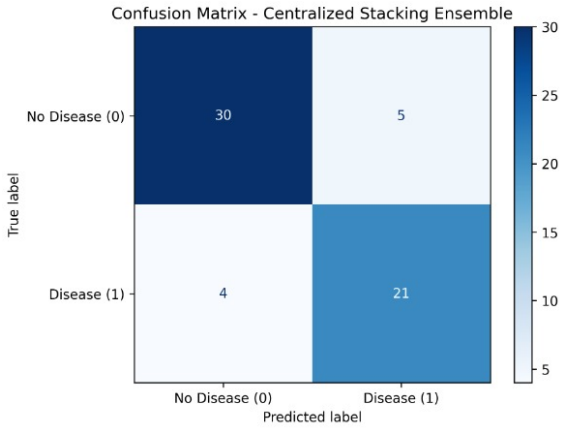

Fig. 4. Confusion Matrix for Centralized Power Boost Ensemble

false positives (5) and false negatives (4). These low error counts stem from the ensemble's crossv validation process, where each base learner was trained on 90% of data and validated on 10%, creating robust out-of-fold predictions. The false negatives cases where disease was missed are primarily samples with borderline values in features such as `oldpeak` and `ca`, which tend to overlap between classes. Conversely, most false positives originate from mildly elevated cholesterol or chest pain cases that resemble disease patterns in the training set.

*3) **Analysis of Loss Metrics and Federated Behaviour**:* Table II shows that the centralized and federated models achieve low Binary Cross Entropy and Hinge losses, indicating good separation between classes. The higher Ridge loss is expected because of the regularization applied to the model weights. The federated setup remains stable even with data spread across four different clients, supported by identical MLP architectures and matched hyperparameters on each site. FedAvg weighting means clients with more samples have a stronger influence on the global update, which explains the higher recall for the disease class in Cleveland and Hungarian datasets where positive cases are more common.

*4) **Federated Global Model Performance** :* In Table III, the federated model achieves balanced F1-scores across both classes despite data being distributed across four heterogeneous clients. This performance stability is strongly influenced by the use of identical MLP architectures and synchronized hyperparameters three local epochs, batch size 16, and Adam with a learning rate of 10^{-3}. The relatively higher recall (0.88) for the disease class indicates that the model places more weight on disease related patterns during FedAvg aggregation. This happens because two hospitals (Cleveland and Hungarian) contain higher proportions of positive cases, and FedAvg weights their updates more heavily through the n_k/N term.

*5) **Federated Confusion Matrix Behavior**:* The model properly detects the majority of illness cases (true positives) from Fig. 5, although it generates a somewhat higher number of false positives. When class imbalance varies among clients, this behavior is typical in federated learning. Switzerland has more sickness cases, but VA Long Beach has a greater percentage of healthy samples. Aggregated gradients that highlight clients with a high percentage of positive cases help the MLP develop a more disease-sensitive border. In addition to contributing to an increase in false alarms, this explains the greater recall.

*6) **ROC–AUC Interpretation**:* The ROC curve in Fig. 6 demonstrates excellent discrimination across thresholds. The smooth curvature and large

TABLE II
COMPUTED LOSS FUNCTION VALUES

Loss Function	Centralised Value	Federated Value
Ridge Loss	0.6240	0.4185
Binary Cross-Entropy Loss	0.4834	0.4185
Hinge Loss	0.4635	0.4874

TABLE III
PERFORMANCE METRICS FOR FEDERATED GLOBAL MODEL

Metric	Class 0: No Illness	Class 1 : Illness	Overall
Precision	0.83	0.79	0.81
Recall	0.71	0.88	0.80
F1-Score	0.77	0.83	0.80
Accuracy	**0.8054**		
ROC–AUC	**0.9009**		

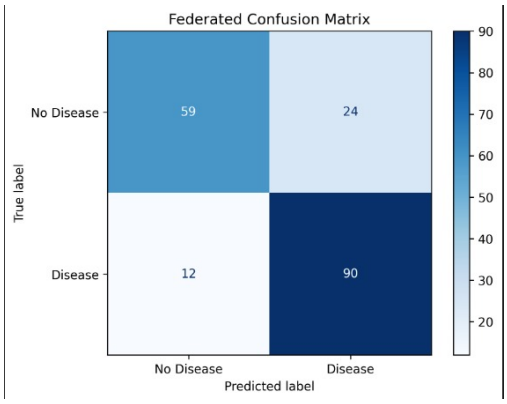

Fig. 5. Confusion Matrix for Federated Global Model

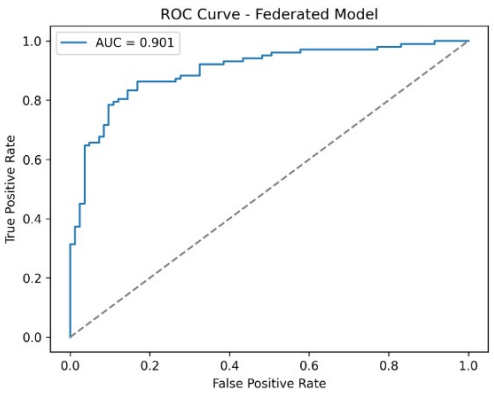

Fig. 6. ROC–AUC Curve for Federated Global Model

AUC are a direct consequence of FedAvg averaging over multiple local MLPs, which smooths decision boundaries by combining gradients from heterogeneous distributions. This reduces overfitting to any single institution and enables more consistent separation of classes.

B. Qualitative Analysis and Explainability

To complement the quantitative performance of the proposed centralized and federated models, a detailed interpretability analysis was conducted using SHAP. This provides a principled game-theoretic framework to evaluate how each input feature contributes to the model output. For a prediction $f(x)$, the SHAP value of feature j is defined in 6.

$$\phi_j(x) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \cdot \left[f(S \cup \{j\}) - f(S) \right] \tag{6}$$

where, F denotes the set of all features. Here, $\phi_j(x)$ measures the marginal contribution of feature j across all possible feature coalitions, thereby ensuring a consistent and locally accurate attribution of predictions. Fig. 7 and 8 show how the centralized stacking model assigns importance to the 14 clinical features. The SHAP summary plot highlights both the strength and direction of each feature's effect on the prediction. Key variables such as ca, cp, thal, and oldpeak stand out as the most influential, which is consistent with clinical understanding that vessel blockage, chest pain patterns, thalassemia markers, and ST depression are strong indicators of heart disease.

IV. Conclusion

This work presented a unified framework for heart disease prediction by combining a centralized stacking ensemble with a privacy-preserving federated learning architecture. The Power Boost Ensemble demonstrated strong generalization by integrating diverse base learners through a Ridge meta-learner, while the federated MLP enabled collaborative training across multiple hospitals without sharing raw patient data. Both approaches showed reliable performance and produced clinically meaningful explanations through SHAP analysis, highlighting features that align with established medical knowledge. Together, these findings emphasize the potential of combining ensemble learning, federated training, and explainability to build trustworthy and secure decision-support systems for healthcare applications.

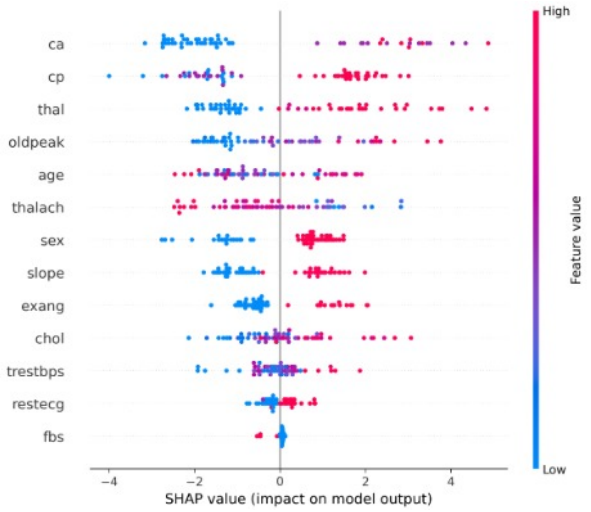

Fig. 7. SHAP summary plot showing local feature contributions for heart disease classification.

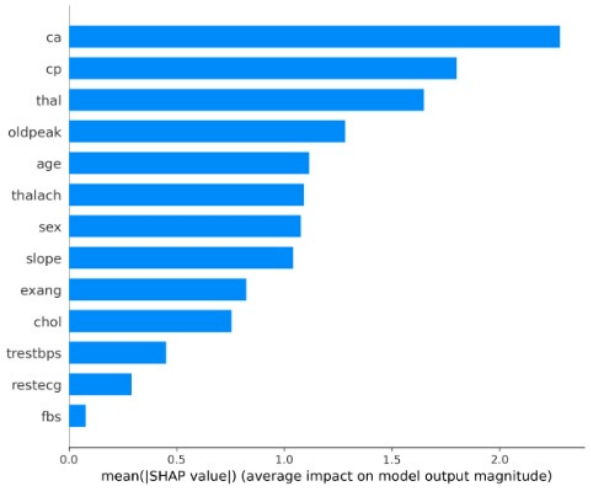

Fig. 8. Global feature importance ranked by mean absolute SHAP values.

REFERENCES

[1] S. Bebortta, S. S. Tripathy, S. Basheer, and C. L. Chowdhary, "Fedehr: A federated learning approach towards the prediction of heart diseases in iot-based electronic health records," *Diagnostics*, vol. 13, no. 20, p. 3166, 2023.

[2] R. Singh and P. Sharma, "Stacked ensemble and shap-driven analysis for heart disease prediction," *Expert Systems with Applications*, vol. 250, p. 124908, 2025.

[3] Y. Zhang, J. Chen, and T. Liu, "Interpretable ensemble-based federated learning for cardiovascular risk assessment," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 2, pp. 345–356, 2024.

[4] T. Rahman and M. Ali, "Explainable ai in healthcare: Enhancing trust in clinical predictions," *Artificial Intelligence in Medicine*, 2024.

[5] Y. Chen and H. Liu, "Trustworthy ai models for clinical decision support systems," *Nature Digital Medicine*, 2024.

[6] A. Gupta and T. Kadian, "Hybrid ensemble frameworks for medical disease prediction: A comparative perspective," *Biomedical Signal Processing and Control*, vol. 89, p. 105023, 2024.

[7] S. Kumar, R. Verma, and P. Chauhan, "A privacy-preserving explainable federated model for multi-hospital cardiac diagnosis," *Computers in Biology and Medicine*, vol. 157, p. 107653, 2025.

[8] R. Detrano *et al.*, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American Journal of Cardiology*, vol. 64, no. 5, pp. 304–310, 1989.

[9] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart disease dataset," UCI Machine Learning Repository, 1988. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease

[10] M. Nafea and A. Yener, "Privacy-preserving federated ensemble learning for personalized healthcare," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 10 122–10 135, 2024.

[11] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 54. PMLR, 2017, pp. 1273–1282, also available as arXiv:1702.02619. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html

[12] H. Yoon, D. Lee, and J. Kim, "Fedstack-xai: Federated stacked learning with explainable insights for clinical prediction," *IEEE Access*, vol. 13, pp. 4511–4524, 2025.

[13] S. Roy, K. Jain, and S. Dey, "Federated learning with explainable insights for cardiac disorder detection across hospitals," *Computers in Biology and Medicine*, vol. 157, p. 107699, 2025.

[14] J. Opitz, "A closer look at classification evaluation metrics and a critical reflection of common evaluation practice," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 1–25, 2024.

[15] O. Rainio, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, p. 56706, 2024.

[16] M. R. Salmanpour, M. Alizadeh, G. Mousavi, S. Sadeghi, S. Amiri, M. Oveisi, A. Rahmim, and I. Hacihaliloglu, "Machine learning evaluation metric discrepancies across programming languages and their components: Need for standardization," *arXiv preprint arXiv:2411.12032*, 2024.

[17] M. Owusu-Adjei *et al.*, "A systematic review of prediction accuracy as an evaluation metric in clinical modelling," *medRxiv*, 2023, preprint.