AIMS Mathematics, 10(3): 7290–7318.
DOI: 10.3934/math.2025334
Received: 08 February 2025
Revised: 17 March 2025
Accepted: 24 March 2025
Published: 28 March 2025

http://www.aimspress.com/journal/Math

Research article

Federated and ensemble learning framework with optimized feature selection for heart disease detection

Olfa Hrizi[1], Karim Gasmi[1], Abdulrahman Alyami[2,*], Adel Alkhalil[3], Ibrahim Alrashdi[1], Ali Alqazzaz[4], Lassaad Ben Ammar[5], Manel Mrabet[5], Alameen E.M. Abdalrahman[1] and Samia Yahyaoui[6]

[1] Department of Computer Science, College of Computer and Information Sciences, Jouf University, Sakaka 72388, Saudi Arabia
[2] Department of Information Systems, College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia
[3] Department of Software Engineering, College of Computer Science and Engineering, University of Hail, Hail, 81481 Saudi Arabia
[4] College of Computing and Information Technology, University of Bisha, Bisha 61922, Saudi Arabia
[5] Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia
[6] Department of Physics, College of Science, Jouf University, Sakaka, Aljouf 72341, Saudi Arabia

* **Correspondence:** Email: am.yami@ju.edu.sa.

Abstract: Predictive models for early identification of heart disease must be precise and efficient because it is a major worldwide health concern. To improve classification performance while protecting data privacy, this study investigated a combined method that uses ensemble learning, feature selection, and federated learning (FL). The ensemble-based approaches proved the most predictive after testing several different machine learning (ML) models, including random forests, the light gradient boosting machine, support vector machines, k-nearest neighbors, convolutional neural networks, and long short-term memory. We used particle swarm optimization (PSO) for feature selection, which optimized the most relevant features in conjunction with voting and stacking approaches to further increase the model's performance. In addition, federated learning was implemented to allow decentralized training while preserving sensitive medical data. The results highlight the effectiveness of combining these techniques in the detection of heart disease, providing a scalable and privacy-preserving solution for real-world healthcare applications. Two benchmark datasets were used to validate the proposed approach, ensuring the reliability and generalizability of the findings. Furthermore, we used four performance metrics, namely accuracy, precision, recall, and F1score, to evaluate the selected models. Finally, federated learning was included to handle privacy issues and guarantee safe access to private medical data. This distributed method allows model training without centralizing patient data, so it is compatible with strict data privacy rules. With up

to 95% precision, our method shows a notable increase in prediction accuracy according to the testing results. This work offers a strong, scalable, and safe solution for the early identification of cardiovascular diseases by combining ensemble learning, feature selection, and federated learning, opening the way for more general uses in medical diagnostics.

Keywords: heart disease detection; machine learning; federated learning; feature selection; optimization
Mathematics Subject Classification: 62H30, 68T05, 92C50, 68U35, 90C59

1. Introduction

Cardiovascular diseases (CVD) remain the leading cause of mortality worldwide, accounting for approximately 17.9 million deaths annually, according to the World Health Organization (WHO) [1]. Early and accurate detection of heart disease is crucial for reducing mortality rates and improving patient outcomes. However, traditional diagnostic methods, such as electrocardiograms (ECG) and clinical tests, often rely on manual interpretation of one-dimensional data, leading to limited accuracy and high variability in diagnosis. Furthermore, these conventional methods lack the ability to integrate diverse and complex patient data for improved decision-making.

Machine learning (ML) has emerged as a promising tool to improve heart disease detection by leveraging large datasets to identify hidden patterns in clinical characteristics [2]. Various ML techniques have been explored, including decision trees, support vector machines (SVM), random forests, and deep learning architectures such as convolutional neural networks (CNN) and long-short-term memory (LSTM) networks. Although these approaches have shown potential in predictive accuracy, they suffer from limitations such as overfitting, data set bias, and the inability to generalize effectively between diverse populations [3–5]. Furthermore, existing ML models often require centralized data storage, raising significant privacy concerns in medical applications.

Several studies have investigated ML techniques for the diagnosis of heart disease. For example, Chaithra and Madhu [3] analyzed transthoracic echocardiography data using artificial neural networks (ANN), J48, Naive Bayes (NB), and decision trees (DT), finding that ANN yielded superior results. Kipp et al. [4] reviewed the role of artificial intelligence (AI) in cardiology and emphasized the effectiveness of deep learning (DL) models in improving diagnostic precision. Furthermore, Sarangam Kodati and Vivekanandam [5] proposed a predictive system based on data mining tools such as Waikato Environment for Knowledge Analysis (WEKA) and Orange, utilizing algorithms such as SVM, NB, and KNN to extract meaningful insights from medical data.

Although these individual models have demonstrated high accuracy, their performance often depends on specific data sets and parameter configurations, leading to variability in the results. Consequently, ensemble learning, which combines multiple models to enhance robustness and predictive power, has emerged as a promising approach to improving heart disease detection. Studies by Shalet et al. [6] and Uyar and lhan [7] have shown that ensemble methods can significantly improve classification accuracy by leveraging the strengths of multiple ML models.

To address these challenges, our study proposes a federated and integrated learning framework with optimized feature selection for heart disease detection. The novelty of our approach lies in integrating

ensemble learning techniques with particle swarm optimization (PSO) to select features while using federated learning (FL) to ensure data privacy. Our methodology consists of three key components.

(1) Ensemble learning: We combine multiple ML classifiers (e.g., random forest, the light gradient boosting machine (LightGBM), and CNN) to enhance the model's robustness and accuracy.
(2) Feature selection optimization: We employ PSO to identify the most relevant predictive features, reducing model complexity while maintaining high accuracy.
(3) Federated learning integration: To mitigate privacy concerns, we adopt FL, which enables decentralized training across multiple healthcare institutions without sharing raw patient data.
(4) By addressing the shortcomings of standalone ML models and centralized learning approaches, our framework provides an efficient, scalable, and privacy-preserving solution for heart disease detection. This research contributes to the growing field of AI-driven healthcare care by enhancing predictive accuracy while safeguarding sensitive patient information.

The remainder of this paper is organized as follows. Section 2 reviews related work on heart disease detection using machine learning. Section 3 describes our proposed methodology, including ensemble learning, feature selection, and federated learning. Section 4 details the experimental setup and datasets. Section 5 presents the results and analysis, followed by a conclusion and future directions in Section 6.

2. Related work

Many studies have investigated the detection and classification of cardiovascular disease (CVD) using machine learning and deep learning methods, applied to various datasets. These papers stress several approaches and techniques, as well as algorithms.

Combining the decision tree and adaptive boosting (AdaBoost) algorithms, Khader Basha et al. [8] create a hybrid model using the Framingham Heart Laboratory dataset, a subset of the Framingham Heart Study (FHS). This data set, which focuses on genetic and environmental factors in cardiovascular disease, was divided into 70% training and 30% tests, containing 16 characteristics. Putting emphasis on the future potential of AI, Kipp W. Johnson et al. [4] summarized the applications of AI and ML in cardiology. Techniques like deep learning (DL) and neural networks (NN) have shown promising results in the diagnosis of cardiovascular diseases. Although these methods require specialized knowledge, they have the potential to transform the field and become more accessible in the near future. Using stochastic gradient descent (SGD), decision trees, and random forest algorithms, Jahed et al. [9] analyzed the effects of stratifying a Kaggle dataset by gender and race. They found that stratification improved accuracy significantly, highlighting the importance of customized data analysis for cardiovascular prediction.

Based on four classification algorithms, random forest, decision trees, logistic regression, and naive bayes, Rajdhan et al. [10] developed a system. The data were divided into two subsets for training and testing, and performance was assessed using a confusion matrix. Random forest achieved the highest accuracy at 90.16%. Das et al. [11] extensively investigated heart disease detection using various machine learning techniques, including XGBoost, bagging, random forest, decision trees, K-nearest neighbors (KNN) and naive bayes. Their study utilized the Kaggle dataset "Key Indicators of Heart Disease" [12], featuring 319,795 cases and more than 300 features. The preprocessing included data cleaning, duplicate removal, and categorical variable transformation, with the models

being evaluated based on precision, sensitivity, precision, the F1 score, and area under the curve (AUC). To identify cardiovascular diseases, Chopra et al. [13] employed principal component analysis (PCA) to reduce the dimensionality of the data from the Cleveland dataset, which contains 303 cases and 14 features. Testing models such as random forest, KNN, decision trees, and naive bayes, the use of PCA significantly improved detection accuracy. To address cardiac rhythm classification, Li et al. [14] employed a deep residual network achieving an impressive accuracy of 99.06%. Similarly, Marinho et al. [15] combined NB, SVM, and the optimum-path forest (OPF), reaching an accuracy of 94.30%. Pandya et al. [16], using the Heart-200 and PhysioNet databases, investigated the acoustic events of heart rhythms. Their InfusedHeart methodology outperformed other models, such as CNN, LSTM, and RNN, showing remarkable performance in identifying auditory abnormalities.

In another effort, Mohan et al. [17] introduced the hybrid random forest linear model (HRFLM). This model integrates decision tree partitioning, error minimization, discriminative feature extraction, and final classification, demonstrating robust performance across multiple datasets. Focusing on cloud security for medical data, Arun R. and N. Deepa [18] utilized naive bayes (NB) and the advanced encryption standard (AES) for their methods. The proposed encryption and re-encryption techniques improved privacy while efficiently handling sensitive patient data. In the context of reinforcement learning, Prasanna et al. [19] used the Cleveland dataset and focused on features such as the resting blood pressure (trestbps), cholesterol, and age. Their Q-learning-based model surpassed traditional machine-learning methods by associating state_action pairs with positive or negative rewards. Bagavathy et al. [20] compared the K-means clustering and MapReduce techniques to identify heart diseases. By integrating MapReduce into distributed systems, they achieved higher accuracy due to its dynamic linear scaling, despite batch processing-induced latency.

To address data imbalances, Abdellatif et al. [21] applied the synthetic minority oversampling (SMOTE) technique. By combining SMOTE with the extra trees method and hyperband hyperparameter tuning, they achieved improved accuracy. Their process included rigorous data cleaning, normalization, and balanced evaluation metrics. Uyar and İlhan [7] proposed a prediction model combining the genetic algorithm (GA) and recurrent fuzzy neural networks (RFNN). Tested on a dataset of 297 cases, the model achieved a high accuracy of 97,78%, emphasizing the importance of metrics such as root mean square error (RMSE) and F-scores.

In their investigation, Ramesh et al. [22] analyzed a Kaggle dataset, finding that 55% of the samples had CVD while 45% did not. Correlation analysis revealed strong relationships, with gradient boost and random forest outperforming other techniques for features like chest pain type and maximum heart rate achieved (MaxHR). Shail et al. [23] assessed machine learning methods, including random forest, decision trees, KNN, and logistic regression, in diverse datasets such as Framingham and UCI. Their findings confirm that logistic regression consistently delivered superior performance. Ali et al. [24] design an expert system using two SVM models optimized using the hybrid grid search algorithm (HGSA). The system outperformed traditional techniques on six evaluation metrics, including AUC and sensitivity. Mahmud et al. [25] explored ensemble learning methods, including bagging, voting, and stacking. With base classifiers such as SVM, decision trees, random forest, and XGBoost as the meta-classifier, they achieved superior results compared with standalone methods. Tama et al. [26] developed a three-phase system for CVD detection. This system integrated XGBoost, gradient boosting machine (GBM) and Random Forest in a two-level set-up and used PSO for feature selection, yielding remarkable validation results on datasets like Cleveland and

Z-Alizadeh. Lastly, Yildirim et al. [27] proposed a deep bidirectional long short-term memory with weighted sum (DBLSTM-WS) model, achieving an extraordinary accuracy of 99.39%. Their use of a wavelet-based layer significantly improved performance, demonstrating the potential of advanced deep learning methods in cardiac diagnosis. Fitriyani et al. [28] designed a decision support system (DSS) for early diagnosis using XGBoost and structured clinical data. Their model delivers high predictive accuracy, demonstrating its utility in assisting clinical decision-making processes.

Using WEKA tools, Devansh Shah et al. [29] implemented supervised learning methods to predict the risks of heart disease. Four categorization models were evaluated: NB, KNN, random forest (RF), and decision tree (DT). Among these, KNN achieved the highest accuracy, highlighting its suitability for such predictive tasks. Using Svetlana Ulianova's CVD dataset, which has 12 characteristics and one target variable, Jin et al. [30] implemented tree-based feature selection methods, including the extra tree classifier and recursive feature elimination (RFE). Tree-based selection evaluated the relevance of the features using the Gini index, while RFE removed the less important features iteratively. Among the models tested, namely random forest, SVM, KNN, and neural networks, neural networks demonstrated the best performance. Furthermore, Shukla et al. [31] focused on data pre-processing by combining a genetic algorithm with recursive feature elimination (GARFE) to select the most relevant features [32]. Missing values were addressed using multivariate imputation by chain equations (MICE). The study used standard normalization and synthetic minority oversampling (SMOTE) to balance classes. Their analysis revealed maximum performance with Logistic Regression and random forest among other algorithms such as naive bayes, SVM, and AdaBoost.

Before employing machine learning methods, Varshini et al. [33] applied the relief feature selection approach and principal component analysis (PCA). PCA identified data patterns, while relief selection retained the most pertinent features. Using the Heart Disease dataset with 11 clinical features and 1,190 cases, random forest outperformed other algorithms. Through a web application, Saranya et al. [34] proposed a cost-effective and time-efficient method to predict heart diseases. After data pre-processing from a hospital in Coimbatore, random forest and KNN achieved accuracies of 100% and 91.36%, respectively. Furthermore, an ensemble model integrating logistic regression demonstrated accuracies of 98.77% and 95.06% with and without logistic regression, respectively.

In this context, Manpreet Singh et al. [35] introduced a heart disease prediction system (HDPS) using structural equation modeling (SEM) and a fuzzy cognitive map (FCM). Drawing from the Canadian Community Health Survey (CCHS), the system used 80% of the data for training and 20% for testing. Despite requiring considerable computation time, this approach achieved 74% accuracy. Sen et al. [36] employed soft voting in an ensemble learning method where the final class label was derived from the averaged probabilities of several models. Their approach incorporated models such as LightGBM, XGBoost, random forest, multilayer perceptron (MLP), gaussian NB, and CatBoost. By aggregating data sets from the UCI Stalog group and the UCI Heart Disease group, they minimized false negatives, a key factor in cardiovascular diagnosis. Furthermore, Gola et al. [37] presented a novel feature selection method based on satin bowerbird optimization (SBO). Their pipeline included handling missing data, normalization, and feature weighting with a modified Kalman filter. By integrating backpropagation and min-max normalization, this approach outperformed conventional techniques.

Using WEKA tools, Jaymin Patel et al. [38] analyzed decision tree-based methods, including J48, the logistic model tree, and the random forest. Although the system effectively identified hidden

patterns within large datasets, its scalability and accuracy required further improvements. Pandita et al. [39] developed a web application to predict heart disease using five machine learning methods. Designed with HTML/CSS and Flask, the tool allowed users to input medical data for risk evaluation. KNN achieved the best accuracy (89.06%), followed by logistic regression (84.38%). Finally, Akella et al. [40] demonstrated that neural networks achieved the highest precision (93.03%) and recall (93.8%) among six models tested in the UCI dataset. The results indicated minimal false negatives, underscoring the accuracy of neural networks in the prediction of heart disease. Random forest, decision trees, KNN, and logistic regression were tested across multiple datasets by Shail et al. [23], including the Framingham, Cleveland, and UCI datasets. Logistic regression consistently outperformed the other models, validating its robustness in the prediction of cardiovascular disease.

This detailed review of the current work emphasizes the variety of approaches used for CVD detection, whether conventional methodologies, ensemble models, or hybrid methods. These analyses also underline the need for feature selection, imbalance control, and algorithm modification to improve model's performance.

3. Proposed model for heart detection

In this article, we propose an innovative approach to classify CVD using machine learning models embedded in an ensemble learning framework. To optimize the efficiency and accuracy of our model, we implemented a feature selection technique based on the PSO algorithm. This approach reduces the dataset's dimensionality by retaining only the most relevant and discriminative features, thereby enhancing the classification model's performance and accuracy. Additionally, we incorporate federated learning in the classification stage to ensure the privacy and security of sensitive medical data. Federated learning enables decentralized model training, where the data remain distributed across local nodes, mitigating the risks associated with centralized data storage while maintaining high performance. Figure 1 presents all the steps of our approach.

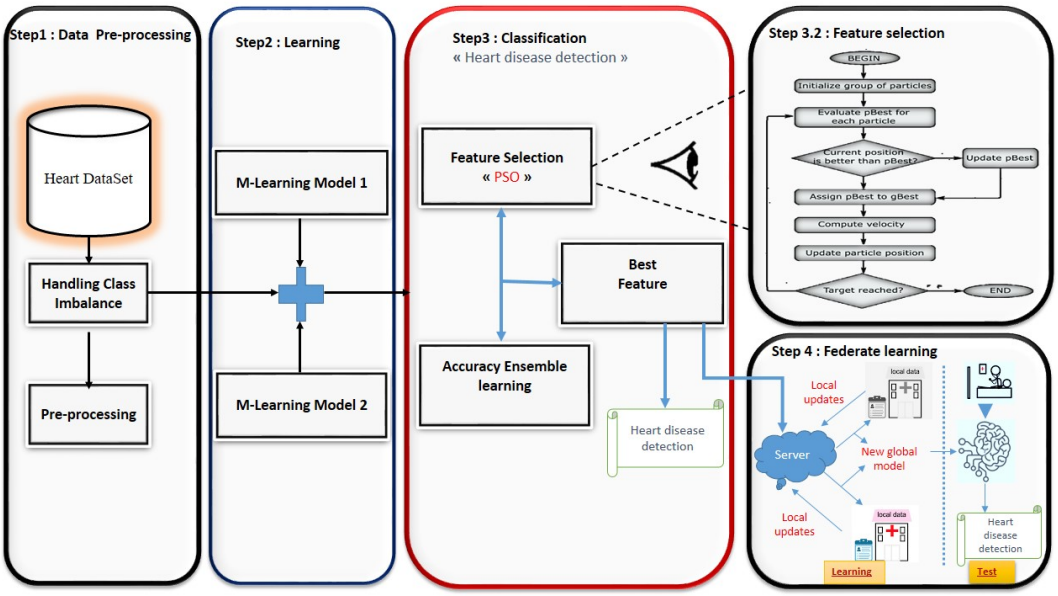

Figure 1. Proposed model for heart disease detection.

Algorithm 1 Proposed approach for heart disease detection

Input: Dataset $\mathcal{D}$ with features $F = \{f_1, f_2, \ldots, f_d\}$
Output: Heart disease detection $\mathcal{H}$
Steps:
procedure PREPROCESSING
    Clean the data: Handle missing values, remove duplicates, and rectify outliers
    Transform the data: Convert features into suitable formats for analysis
    Normalize the data: scale features to a standard range
    Balance the data: address class imbalances using techniques like SMOTE
end procedure
procedure FEATURE SELECTION
    Initialize the PSO algorithm
    for each particle p_i in the swarm **do**
        Select a subset of features $F_i \subseteq F$
        Compute the fitness function: Fitness(F_i) = Accuracy$(\mathcal{M}, F_i)$
    end for
    Select the optimal feature subset F^*
end procedure
procedure ENSEMBLE MODEL TRAINING
    Train base models: SVM, KNN, LightGBM, and CNN on F^*
    Combine predictions using ensemble techniques:
    - Hard voting: Majority class prediction
    - Soft voting: Weighted average of probabilities
    - Weighted average voting: Assign weights to classifiers
    - Stacking: Use the meta-classifier for final predictions
end procedure
procedure FEDERATED LEARNING INTEGRATION
    Initialize global model w^0
    for each node i **do**
        Train local model w_i^t on D_i using selected features F^*
        Compute model updates $\Delta w_i^t = w_i^{t+1} - w_i^t$
    end for
    Aggregate updates: $w^{t+1} = \frac{\sum_{i=1}^{N} |D_i| w_i^t}{\sum_{i=1}^{N} |D_i|}$
    Distribute updated global model w^{t+1} to all nodes
end procedure
Return: Heart disease detection $\mathcal{H}$

3.1. Dataset

The first cardiovascular disease dataset used in this paper comprises 12 variables and has 70,000 instances that were acquired via medical examinations. The first 11 variables are considered input characteristics, while the 12th variable is considered the output characteristics, indicating whether or not cardiovascular disease is present. The fact that this data set has a significant number of duplicate

values and extreme outliers is something that should be noted. Consequently, during the preprocessing stage, duplicated values and cases with extreme outliers were eliminated, resulting in a reduction in the total number of records to 62,267. In addition, the attributed age was taken from days and translated into years. For the purpose of conducting a more thorough analysis, the systolic and diastolic blood pressure readings were converted from numerical to nominal using the normal range established by the American Heart Association.

The second data set contains 1190 instances that were obtained by combining five original datasets across 11 shared variables and one attribute that served as a predictor to determine whether or not a patient had cardiovascular disease. Five original datasets were incorporated into the creation of this data set. These datasets include: Cleveland (303 samples), Hungarian (294 samples), Switzerland (123 samples), Long Beach, Virginia (200 samples), and Statlog data sets (270 samples). Additionally, this dataset contains values that were missing or duplicated, which were eliminated during the preprocessing step, reducing the total number of records to 918.

3.2. Dataset bias and class imbalance

The data sets used in this study provide valuable information on heart disease detection; however, it is important to acknowledge potential biases, particularly in terms of class imbalance and demographic skewness, which can affect the generalization of the model. Dataset 1 has a nearly balanced class distribution, with 50.03% of samples labeled as nondisease and 49.97% as heart disease cases, ensuring no strong class imbalance. However, Dataset 2 exhibits a mild imbalance, with 52.86% of cases diagnosed with heart disease and 47.14% classified as non-disease cases. Although this imbalance is not extreme, it may introduce a slight prediction bias toward the majority class.

Both sets of data exhibit strong gender bias, in addition to the class imbalance that is present. Dataset 1 has a gender imbalance of 65.04% percent male and 34.96% percent female participants. However, Dataset 2 has an even more lopsided gender ratio, with 76.39% percent male participants and only 23.61% percent female participants. A conclusion that can be drawn from this is that models that have been trained on these datasets might be more reliable for male patients, but might be less accurate for female patients. Furthermore, these data sets do not include clear demographic features such as age distributions, ethnicity, or socioeconomic status. These attributes are critical in determining whether or not a model is fair when applied to real-world healthcare applications. It is necessary to conduct additional testing on populations that are more diverse and representative to determine whether or not these demographic imbalances have an effect on the generalizability of models that were trained on these datasets.

3.3. Data preprocessing

Data preprocessing is a crucial phase in machine learning, involving several tasks designed to convert the information into a suitable format for efficient analysis and effective training. Initially, we manage absent data, eliminate redundancies, correct errors, and confront outliers. This constitutes a data cleansing phase that includes identifying and correcting errors or inconsistencies. This involves the cleansing and conversion of data into optimal formats for analysis, enhancing the efficacy of machine learning models by rendering the data more manageable and interpretable. The subsequent

step involves the correcting of class disparities. This guarantees that the model does not preferentially represent the majority class and can reliably generate predictions for all classes. Finally, we standardize the data by scaling the characteristics to a uniform range. This mitigates the risk of any characteristic exerting an excessively large impact on the learning process due to scale discrepancies. The preparation stages ensure the data's quality and efficacy for machine learning applications.

3.4. Ensemble learning

Combining several classifiers into an ensemble has successfully produced strong high-performance prediction models. Using the strengths of individual classifiers, an ensemble technique improves input instance-based decision-making. In this work, we used several basic classifiers in an ensemble approach to identify cardiac disorders. Following an extensive review of many machine learning algorithms, including decision trees, support vector machines (SVM), logistic regression, KNN, AdaBoost, extra trees, random forest, a gradient boosting algorithm (LightGBM), and deep learning models such as convolutional neural network (CNN) and long short-term memory (LSTM), these classifiers were chosen.

3.4.1. Support vector machines

SVMs are supervised learning systems widely applied to classification and regression tasks. Their accuracy in handling high-dimensional data and generating correct results is especially well known. SVMs work by determining the ideal hyperplane in a N-dimensional feature space to divide the data points into their respective classes.

3.4.2. Decision trees

Random forest is an ensemble learning technique that constructs many decision trees during training and aggregates their output to improve classification or regression performance. It works by choosing random subsets of data and features to reduce overfitting and improve model generalization. Using feature significance analysis, this method provides great accuracy and interpretability, is robust to noise, and performs effectively on large datasets.

3.5. Logistic regression

Binary and multiclassification problems use logistic regression as a statistical model. The logistic function applied to a linear combination of input features approximates the probability that a given input belongs to a particular class. When probabilities are needed and when linearly separable data are involved, this method is appropriate, since it is computationally efficient, interpretable, and straightforward to implement.

3.5.1. Long short-term memory

Designed to manage sequential data by learning long short-term dependencies, long-term memory (LSTM) is a kind of recurrent neural network (RNN). It selectively stores and updates data using gated memory cells (input, forget, and output), thus avoiding the vanishing gradient issue typical in conventional RNNs. LSTMs is quite successful for jobs including speech recognition, natural language processing, and time series analysis.

3.5.2. K-nearest neighbors

Often used in both classification and regression, KNN is a non-parametric, instance-based learning technique. Its simplicity resides in its method of classification, whereby the k-nearest data points guide an input instance to the class most like it. A preset distance metric, say Euclidean distance, guides the choice of neighbors.

3.5.3. Light gradient boosting machine

Developed by Microsoft in 2017, the LightGBM is a highly performing gradient boosting tool. Based on decision tree techniques, LightGBM excels in ranking, classification, and regression problems. Large-scale machine learning applications would find it perfect, since it is known for its low memory usage and quick training periods, which help it to efficiently manage large datasets.

3.5.4. Convolutional neural networks

Convolutional neural networks (CNNs) reflect a deep learning architecture. They use convolutional operations in their layers and have been used in natural language processing (NLP) applications, but their main applications are picture classification and recognition. CNN architectures comprise fully connected layers for classification, pooling layers for dimensionality reduction, and convolutional layers for feature extraction. CNNs are very strong for complicated data analysis, as convolutional layers can detect spatial and hierarchical aspects.

3.5.5. Hard voting

In the ensemble learning method known as " hard voting," each base classifier votes for a predicted class, and the majority vote decides the final prediction. It considers every classifier equally independent of their respective performance. Implementing this approach is easy and works best when the underlying classifiers are varied and accurate. It is appropriate for classification problems when most classifier agreements are sought. In hard voting, the final prediction $\hat{y}$ is determined by taking the majority vote among the predictions of the n classifiers as follows:

$$\hat{y} = \text{mode}\left(\{y_1, y_2, \ldots, y_n\}\right)$$

where: - $\hat{y}$: is the final predicted class; - $y_1, y_2, \ldots, y_n$, is the predicted class labels of the n classifiers, and - mode is the function that returns the class with the highest frequency among $\{y_1, y_2, \ldots, y_n\}$.

3.5.6. Soft voting

Soft voting compiles the expected probabilities of each base classifier for every class and chooses the final prediction on the basis of the average probability. Unlike hard voting, it takes individual classifiers' confidence into account, thereby providing greater weight to those classifiers whose predictions are more definite. Particularly in cases where the classifiers are probabilistic in character, soft voting helps to increase accuracy. In soft voting, the final prediction $\hat{y}$ is based on the class k with the highest average probability as follows:

$$\hat{y} = \arg\max_k \left(\frac{1}{n} \sum_{i=1}^{n} P_i(k) \right)$$

where: - $\hat{y}$ is the final predicted class, - $P_i(k)$is the predicted probability of class k by the i-th classifier, - n is the total number of classifiers, - $\arg\max_k$ is the function that identifies the class k with the highest average probability.

3.5.7. Weighted average voting

In weighted average voting, base classifier predictions are aggregated under weights that correspond to their individual dependability or performance. High-weight classifiers affect the final prediction more strongly. Putting emphasis on the more accurate ones, this method balances the contributions of classifiers and is particularly helpful in ensembles including classifiers with varying strengths. In weighted average voting, the final prediction $\hat{y}$ is determined by the class k with the highest weighted sum of probabilities

$$\hat{y} = \arg\max_k \left(\sum_{i=1}^{n} w_i \cdot P_i(k) \right)$$

where - $\hat{y}$ is the final predicted class, - w_i is the weight assigned to the i-th classifier, - $P_i(k)$ is the predicted probability of class k by the i-th classifier, - n is the total number of classifiers, and - $\arg\max_k$ is the function that identifies the class k with the highest weighted probability.

3.5.8. Stacking

Stacking is a sophisticated ensemble learning technique by which several base classifiers are combined in the training of a metaclassifier on their output. The basis classifiers generate predictions that the meta-classifier, learning how to best combine them, uses as input features. This method uses the advantages of several models to provide better accuracy and flexibility, but must be carefully tuned to avoid overfitting.

In stacking, the metaclassifier h_m is trained on the output of the base classifiers $h_1(x), h_2(x), \ldots, h_n(x)$, and the final prediction $\hat{y}$ is calculated as:

$$\hat{y} = h_m\left(h_1(x), h_2(x), \ldots, h_n(x)\right)$$

where: - $\hat{y}$ is the final predicted class, - $h_1(x), h_2(x), \ldots, h_n(x)$ is the predictions of the n base classifiers for the input x, and - h_m is the metaclassifier that learns to optimally combine the outputs of the base classifier.

3.6. Feature selection

Inspired by the social behavior of flocks of birds or schools of fish, PSO is a metaheuristic optimization method [41]. It is mostly used for feature selection to lower the dimensionality and increase the performance of the model. In this work, we selected the most relevant features by using a PSO applied to two datasets. We used PSO on two datasets: one has 12 features, and the other has 11. The aim was to minimize the number of features chosen for every dataset, while improving the accuracy of the classification model. Reducing the feature set helped us to strike a compromise between computational economy and performance. Each particle in the swarm represents a subset of features

$$X = [x_1, x_2, \ldots, x_d]$$

where

$$x_i \in \{0, 1\} \quad \text{indicates if the } i\text{-th feature is selected or not,}$$
$$d \text{ is the total number of features in the dataset.}$$

Velocity and position updates
The velocity and position updates for each particle are governed by the following equations:
Velocity update:

$$v_{i,j}^{t+1} = \omega v_{i,j}^{t} + c_1 r_1 (p_{i,j} - x_{i,j}) + c_2 r_2 (g_j - x_{i,j}).$$

Position update:

$$x_{i,j}^{t+1} = \begin{cases} 1 & \text{if } \sigma(v_{i,j}^{t+1}) > \text{rand}() \\ 0 & \text{otherwise.} \end{cases}$$

where

$v_{i,j}^{t}$ is the velocity of particle i for feature j at iteration t,
$x_{i,j}^{t}$ is the position of particle i for feature j at iteration t,
ω is the inertia weight,
c_1, c_2 are the cognitive and social coefficients,
r_1, r_2 are random values in $[0, 1]$,
$p_{i,j}$ is the personal best position of particle i for feature j,
g_j is the global best position of the swarm for feature j,
$\sigma(v) = \dfrac{1}{1 + e^{-v}}$ is the sigmoid function,
rand() is a random number in $[0, 1]$.

Fitness function: The fitness function evaluates the subset of characteristics of each particle according to classification accuracy:

$$\text{Fitness}(X) = \text{Accuracy}(X),$$

where

$\text{Fitness}(X)$ is the fitness score for the feature subset X,
$\text{Accuracy}(X)$ is the classification accuracy achieved by the feature subset X.

Optimization aims to maximize accuracy while minimizing the number of selected features.
Optimization objective:

$$\max_{X} \ \text{Fitness}(X) \quad \text{subject to} \quad \sum_{i=1}^{d} x_i \text{ is minimal.}$$

3.7. Federated learning for heart disease detection

The developing paradigm of decentralized machine learning known as federated learning (FL) makes it possible for different clients to train a global model without actually sharing their raw data. The protection of data privacy and security is ensured by this, which makes FL an excellent choice for applications in the healthcare industry, where maintaining patient anonymity is of the utmost importance. FL enables local devices to train models independently, sharing model updates only with a central server for the purpose of aggregation, in contrast to traditional centralized learning, which aggregates data in a single location.

The federated learning process can be described mathematically as follows:

Local model training

Each node i trains a local model w_i^t using its private data set D_i at iteration t. The local objective function is given by:

$$L_i(w) = \frac{1}{|D_i|} \sum_{(x,y)\in D_i} \ell(w; x, y)$$

where

- $L_i(w)$is the local loss function for node i,
- $|D_i|$ is the size of the local dataset D_i,
- $\ell(w; x, y)$ is the loss function (e.g., cross-entropy) for a sample (x, y),
- w is the model parameters.

Model updates

After training, each node computes its update Δw_i^t, which represents the change in its model parameters:

$$\Delta w_i^t = w_i^{t+1} - w_i^t$$

where

- Δw_i^t is the change in model parameters for node i at iteration t,
- w_i^t is the model parameters of node i at iteration t,
- w_i^{t+1} is the updated model parameters of node i in iteration $t + 1$.

Server aggregation

The central server aggregates updates w_i^t from all nodes to compute the global model w^{t+1}. The most common method is federated averaging (FedAvg):

$$w^{t+1} = \frac{\sum_{i=1}^{N} |D_i| w_i^t}{\sum_{i=1}^{N} |D_i|}$$

where

- w^{t+1} is the global model parameters after aggregation,
- N is the total number of nodes,
- $|D_i|$ is the size of the local dataset for node i,
- w_i^t is the model parameters of node i in iteration t.

Global model distribution

The global model w^{t+1} is distributed back to all nodes, where it serves as the starting point for the next training round. The process repeats until convergence.

3.7.1. Impact of FL on communication efficiency and computational cost

As a result of the fact that several clients are required to communicate model updates with a central server, one of the most significant issues that FL faces is the communication overhead. A direct relationship exists between the number of communication rounds and the performance of FL. A lower number of rounds reduces the strain on the network but it slows down convergence. On the other hand, frequent updates improve learning but increase bandwidth utilization. In order to maximize the effectiveness of our communication, we made use of an adaptive aggregation technique. This strategy selectively updates just the most important model parameters, thus minimizing the number of transmissions that are not necessary.

Due to the fact that each node is required to carry out local training prior to taking part in the global update, FL has the additional effect of introducing additional computational constraints on the client side. The computational cost is affected by a number of factors, including the batch size, the number of local epochs, and the complexity of the model. It may be necessary to make careful modifications to the local training process in order to accommodate resource-limited devices, which may result in slower training. Through fine-tuning the FL parameters, we achieved a balance between computational economy and performance in this study. This allowed us to guarantee steady model updates while minimizing the resource overhead.

3.7.2. FL's impact on model convergence and accuracy

Because of the variety of the data among various clients, which has an effect on model convergence, FL training might be difficult to successfully complete. FL nodes train on different local datasets, which can result in potential differences in model updates. This is in contrast to centralized learning, which makes all of the data available throughout the entire process. The federated averaging (FedAvg) algorithm, which aggregates local models in a weighted manner to improve stability of convergence, is the solution that we employed in order to address this problem. The distribution of training data across clients is another element that influences the performance of FL. When local datasets are not identically distributed across nodes, the model may need more rounds of aggregation in order to achieve stable learning. This is because datasets that are not independent and identically distributed (IID) are not identically distributed. In order to answer this question, we investigated the influence that data partitioning schemes and client heterogeneity have on the convergence of FL. In the Results section, a complete analysis of FL's performance is offered. This analysis includes an examination of its influence on the quality of its accuracy and its convergence behavior.

The FL methodology is an effective method for collaborative machine learning because it offers several significant benefits. One of the most important advantages is privacy protection, which is achieved by maintaining data decentralization on local nodes. This prevents sensitive information from being shared or exposed. Furthermore, this is of utmost importance in fields such as healthcare, which are subject to severe regulations regarding data privacy. In addition, FL is very scalable, which makes it possible to collaborate across different nodes, such as hospitals or universities, without the need for data to be centralized. The framework improves efficiency with optimal feature selection. This reduces the computational effort needed for training by focusing on the most relevant features of the task. Furthermore, federated learning facilitates customization, allowing each node to adjust the global model to accommodate its specific local dataset. This improves the performance and effectiveness of

the model with diverse data distributions. Collectively, these advantages make federated learning an appropriate option for conducting extensive collaborative machine-learning endeavors that prioritize privacy considerations.

4. Experimental results and discussion

This section presents the experimental results obtained from the proposed model and evaluates its performance in various classification approaches within the framework of machine learning. The model was implemented using Python and was executed on a system equipped with 16 GB of RAM and an RTX 2060 graphics card. To thoroughly assess the proposed model, we conducted multiple experiments that utilized different deep learning architectures. The design of the model was structured into four distinct scenarios.

(1) Scenario 1: Detection of heart disease using traditional machine learning models.
(2) Scenario 2: Classification of heart disease using hybrid machine learning models.
(3) Scenario 3: Classification of heart disease with an emphasis on feature selection methods.
(4) Scenario 4: Classification of heart disease based on federated learning approaches.

4.1. Evaluation metrics

To assess the performance of our classification approach, we employed standard evaluation metrics, including accuracy, precision, recall, F1 score, and receiver operating characteristic-AUC. Precision measures the accuracy of the model's positive predictions, while recall evaluates the true positive rate. The F1 score combines precision and recall to provide a balanced perspective on the model's ability to handle both false positives and false negatives.

$$Accuracy = TP + TN/(TP + TN + FP + fN), \tag{4.1}$$

$$Precision = TP/(TP + FP), \tag{4.2}$$

$$Recall = TP/(TP + FN)(3), \tag{4.3}$$

$$F1 - score = 2/(1/P + 1/R)(4), \tag{4.4}$$

where TP denotes true positives, FP denotes false positives, P denotes the precision rate, R indicates the recall rate, TPR represents the true positive rate, and FPR signifies the false positive rate.

4.2. Performance comparison of machine learning models for heart disease detection

In this study, we evaluated multiple machine learning classifiers on a heart disease detection dataset, analyzing their performance in terms of their accuracy, precision, recall, and F1 score. The results highlight significant differences in the models' effectiveness, providing insight into their predictive capabilities.

The evaluation of numerous classifiers for the identification of heart disease reveals significant performance discrepancies between the different models. Table 1 presents the results obtained from various classifiers, which include accuracy, precision, recall, and the F1 score. This allows for a thorough comparison of their prediction skills. The classifiers that were examined included random

forest and LightGBM, which had the highest accuracy at 94.54%. Decision trees and KNN were close behind, with accuracies of 90.34% and 88.66%, respectively. The results show that ensemble-based models are generally better than standard classifiers when it comes to prediction power. LightGBM and random forest also demonstrated great precision (94.70% and 93.38%, respectively) and recall (95.42% and 96.95%, respectively), which means that they are very reliable when it comes to reducing false positives while accurately identifying actual cases. KNN had the highest recall rate (93.13%), which indicates that it is very good at identifying cases of heart disease. However, its precision rate (87.14%) was slightly lower, meaning it has a higher false positive rate. Although decision trees had the highest precision (93.55%), this method had a lower recall (88.55%), which means that it is a more conservative model that decreases false alarms but may miss some actual cases.

Table 1. Performance comparison of machine learning models for heart disease detection: Dataset 1.

Model	Accuracy	Precision	Recall	F1 Score
Logistic regression	0. 8613	0. 8712	0. 8712	0. 8745
KNN	0. 8866	0. 8714	0. 9313	0. 9004
SVM	0. 8445	0. 8561	0. 8626	0. 8593
Random forest	0. 9454	0. 9338	0. 9695	0. 9513
AdaBoost	0. 8782	0. 8750	0. 9084	0. 8914
LightGBM	0. 9454	0. 9470	0. 9542	0. 9506
Decision trees	0. 9034	0. 9355	0. 8855	0. 9098
CNN	0. 8361	0. 8485	0. 8550	0. 8517
LSTM	0. 8529	0. 8582	0. 8779	0. 8679

Deep learning methods, such as CNN and LSTM, are often employed for complicated feature extraction. However, they showed lower accuracy (83.61% and 85.29%, respectively) than typical machine learning models. LSTM was able to surpass CNN in all criteria by a small margin by taking advantage of its capacity to capture temporal relationships, although it still fell short of the ensemble models. Ensemble approaches like random forest and LightGBM regularly produced high F1 scores (95.13% and 95.06%, respectively), demonstrating their reliability. Although AdaBoost had good performance (87.82% precision), it was not as good as random forest and LightGBM. This suggests that boosting approaches can enhance performance but do not always outperform bagging-based strategies. The SVM and logistic regression models achieved reasonable performance, with accuracy rates of 84.45% and 86.13%, respectively. Because their recall is lower (85.26% and 87.12%), it is possible that they will have difficulty recognizing all positive cases, making them less suitable for use in high-risk medical situations.

The results shown in Table 2 demonstrate that the ensemble models, especially random forest and LightGBM, are the best classifiers for the detection of cardiac disease. This is because they combine high precision and recall, which makes them appropriate for use in clinical settings. Although deep learning models did not perform better than standard classifiers, their effectiveness could be improved with additional tuning and feature engineering. KNN and ensemble models should be prioritized for applications where it is critical to have a high recall, such as when it is necessary to ensure that all cases of heart disease are diagnosed. However, if minimizing false positives is the most important

factor, decision trees and LightGBM provide excellent results.

Table 2. Performance comparison of machine learning models for heart disease detection with the total ranking score (TRS).

Model	Accuracy	Precision	Recall	F1 Score	TRS
Logistic regression	0.7234	0.7454	0.6798	0.7111	24
KNN	0.6279	0.6316	0.6167	0.6240	36
SVM	0.7284	0.7326	0.7285	0.7273	19
Random forest	0.7278	0.7402	0.7034	0.7213	23
AdaBoost	0.7294	0.7704	0.6550	0.7080	21
LightGBM	0.7397	0.7608	0.7005	0.7294	12
Decision trees	0.6351	0.6338	0.6426	0.6382	28
CNN	0.7326	0.7428	0.7328	0.7299	12
XGBOOST	0.7385	0.7609	0.6969	0.7275	11

Table 2 illustrates the performance of several classifiers in the second dataset, offering information on the generalizability of different models in different data sets. The second dataset exhibits a more equitable performance among classifiers, unlike the previous dataset, where the ensemble models were the most effective. LightGBM achieved the highest accuracy at 73.97%, followed closely by XGBoost at 73.85% and CNN at 73.26%. The results demonstrate that LightGBM remains a valuable model, despite the overall performance metrics being inferior to those of the first data set. Logistic regression and SVM exhibited comparable performance, achieving accuracy rates of 72.34% and 72.84%, respectively. CNN performed better on this data set than on the initial one, indicating its ability to comprehend non-linear patterns in the data.

When you take a closer look at precision and recall values, you can see the trade-offs between models. LightGBM and XGBoost showed excellent precision (76.08% and 76.09%, respectively) but lower recall values (70.05% and 69.69%, respectively), suggesting that they may be more conservative in identifying positive situations. On the other hand, AdaBoost had the best precision (77.04%) but the lowest recall (65.50%), which means that it is very selective but may miss some cases of heart disease. The decision trees and KNN algorithms did not perform well in this dataset, achieving accuracy ratings of 63.51% and 62.79%, respectively. This suggests that they have limited generalizability.

The total ranking score (TRS) evaluates the performance of the model using the accuracy, precision, recall, and the F1 score. A lower TRS means higher performance across all measures. XGBoost had the lowest TRS (11), making it the best model in this evaluation. XGBoost balances precision and recall to perform well in categorization. LightGBM and CNN, both with a TRS of 12, both showed strong predictive power, proving that ensemble-based models (for example, boosting) and deep learning models (CNN) are good for heart disease detection. KNN had the highest TRS (36), suggesting poorer performance in all criteria. KNN may fail to discriminate instances of heart disease due to its sensitivity to feature scaling and the curse of dimensionality. The decision tree has a high TRS (28), confirming its tendency to overfit with limited datasets. The results show that boost-based models (XGBoost, LightGBM) and deep learning models (CNN) can capture complicated medical data patterns and diagnose heart disease better. Ensemble learning methodologies boost model resilience, as shown by XGBoost and LightGBM's success.

When comparing the findings from both datasets, it is clear that ensemble models such as LightGBM and random forest perform well across various datasets, although their effectiveness may change. In the second dataset, CNN and SVM produced more consistent findings, but KNN and decision trees had less generalizability. The findings highlight how important the qualities of a dataset are when it comes to affecting how well a model performs. The first data set produced better results with tree-based ensemble models, while the second data set produced more competitive results across various approaches, including CNN and SVM. These results indicate that there is no one model that is better than all the others, and the choice of model should depend on the data set. In the future, research could investigate model-assembling techniques to combine the characteristics of different methodologies and further increase the prediction accuracy for heart disease detection.

4.3. Evaluation of enhancing heart disease detection through classifier combinations

To improve classification performance, we explored different combinations of models using voting (hard and soft) and stacking techniques. Tables 3 and 4 present the accuracy results for various pairings of classifiers across two datasets, providing information on how ensemble techniques improve predictive capabilities. The results indicate that stacking consistently outperforms hard and soft voting, reinforcing its effectiveness in leveraging multiple models' strengths.

The combination of random forest + decision trees employing stacking achieved the highest precision of 95. 80% for the first dataset, as shown in Table 3. This shows that tree-based models may take advantage of their complementary capabilities when they are effectively integrated. In a similar vein, logistic regression + random forest utilizing stacking achieved an accuracy of 95.38%, demonstrating that a linear model can greatly benefit from a strong ensemble technique. LightGBM performed well in hybrid ensembles, as demonstrated by the high precision of stacking when combined with logistic regression and random forest (94.96% and 94.54%, respectively). Soft voting performed a little better than hard voting, but stacking was still the best option. When AdaBoost was used with other models, it achieved its highest accuracy of 94.96% when combined with LightGBM using stacking, demonstrating its ability to successfully manage weak learners.

Table 3. Evaluation of enhancing heart disease detection through classifier combinations: Dataset 1.

Model	Random Forest			AdaBoost			SVM		
Voting	Hard	Soft	Stacking	Hard	Soft	Stacking	Hard	Soft	Stacking
Logistic regression	0. 9118	0. 9412	0. 9538	0. 8655	0. 8655	0. 8697	0. 8571	0. 8571	0. 8571
Random forest	–	–	–	0. 9202	0. 9496	0. 9538	0. 9034	0. 9286	0. 9538
AdaBoost	–	–	–	–	–	–	0. 8697	0. 8655	0. 8655
SVM	–	–	–	–	–	–	–	–	–
Model	LightGBM			Decision trees					
Voting	Hard	Soft	Stacking	Hard	Soft	Stacking			
Logistic regression	0. 9034	0. 9370	0. 9496	0. 8655	0. 9034	0. 9034			
Random forest	0. 9454	0. 9496	0. 9454	0. 9034	0. 9034	0. 9580			
AdaBoost	0. 9160	0. 9496	0. 9496	0. 8739	0. 9034	0. 9118			
SVM	0. 8992	0. 9412	0. 9496	0. 8571	0. 9034	0. 9034			

Table 4. Evaluation of enhancing heart disease detection through classifier combinations: Dataset 2.

Model	Random forest			AdaBoost			Decision trees		
Voting	Hard	Soft	Stacking	Hard	Soft	Stacking	Hard	Soft	Stacking
Logistic regression	0. 7216	0. 7334	0. 7334	0. 7189	0. 7301	0. 7341	0. 6811	0. 6351	0. 7257
Random forest	–	–	–	0. 7254	0. 7340	0. 7362	0. 6865	0. 6351	0. 7278
AdaBoost	–	–	–	–	–	–	0. 6860	0. 6351	0. 7316
Decision trees	–	–	–	–	–	–	–	–	–
Model	LightGBM			KNN					
Voting	Hard	Soft	Stacking	Hard	Soft	Stacking			
Logistic regression	0. 7276	0. 7377	0. 7389	0. 6704	0. 6829	0. 7242			
Random forest	0. 7346	0. 7364	0. 7383	0. 6695	0. 7010	0. 7284			
AdaBoost	0. 7314	0. 7387	0. 7391	0. 6679	0. 6795	0. 7331			
Decision trees	0. 6915	0. 6351	0. 7404	0. 6286	0. 6390	0. 6556			

The second data set showed a similar pattern, with decision trees + LightGBM utilizing stacking to achieve maximum accuracy (74.04%). This suggests that combining a tree-based model with gradient boosting improves generalization. In a similar vein, AdaBoost + LightGBM using stacking achieved an accuracy of 73.91%, demonstrating the advantages of combining boosting algorithms. When random forest was paired with LightGBM by stacking, the accuracy reached 73.83%. This confirms that ensemble learning improves performance even further. Compared with logistic regression, the best combination was found to be LightGBM employing stacking (73.89%). This suggests that combining a linear model with boosting techniques can help capture complex correlations in the data. However, combinations that included KNN typically produced lower accuracy, indicating that it may not be as successful in hybrid contexts.

Stacking was the most successful ensemble strategy, since it consistently improved classifier performance across both datasets. Tree-based and boosting models (random forest, LightGBM, and AdaBoost) continued to perform well, but their effectiveness was much greater when they were paired with logistic regression and decision trees. The results indicate that stacking hybrid models is a more precise and reliable method of detecting cardiac disease, especially in clinical settings where accurate predictions are essential. Future studies could investigate how to increase the models' performance even more by improving feature selection and hyperparameters inside stacking frameworks.

4.4. Evaluation of optimized heart disease detection using PSO-based feature selection

After evaluating various classifier combinations, we further optimized our model by implementing PSO to select the most significant features, aiming to enhance predictive accuracy while reducing computational complexity. Tables 5 and 6 present the performance results of models trained with different subsets of selected features using PSO for the first and second datasets. The results indicate that optimal feature selection significantly impacts the models' performance across both datasets.

Table 5 shows that the first data set performed the best when nine features were selected. The results were a precision of 95.80%, a precision of 94.81%, a recall of 97.71%, and an F1 score of 96.24%. This indicates that choosing a more specific collection of features improves a model's ability to maintain

a balance between precision and recall. It is interesting to note that employing all available features did not result in any improvement. Both the 10-feature and all-feature models achieved an accuracy of 95.38%, suggesting that some features may add redundancy or noise. The model that was trained with eight features also performed well, achieving an accuracy of 94.54%. This shows that a slightly reduced feature count still retains good predictive potential. It should be mentioned that the model achieved an accuracy of 91.18% when only six features were selected, indicating that the model was still able to classify quite well even with a smaller number of features.

Table 5. Evaluation of optimized heart disease detection using PSO-Based Feature Selection: Dataset 1. Wilcoxon signed-rank test p-values compare each feature selection configuration against the baseline (all features).

Number of features	Accuracy	Precision	Recall	F1 score	p-value
All features	0.9538	0.9478	0.9695	0.9585	-
6	0.9118	0.9231	0.9160	0.9195	0.125
8	0.9454	0.9538	0.9466	0.9502	0.250
9	0.9580	0.9481	0.9771	0.9624	0.125
10	0.9538	0.9478	0.9695	0.9585	-
7	0.9160	0.9512	0.8931	0.9213	0.250

Table 6 shows the best performance for the second dataset. This was reached when nine characteristics were selected, resulting in an accuracy of 74.11%, a precision of 76.13%, a recall of 70.37%, and an F1 score of 73.13%. This result is very similar to the performance of the seven-feature model (73.95% accuracy), suggesting that it is important to use a balanced amount of features to maintain the accuracy of the classification. Interestingly, using all available features resulted in slightly lower accuracy (73.97%). This further supports the theory that feature selection reduces redundancy while boosting a model's generalization. The model with eight selected features performed similarly well (74.06% accuracy), reinforcing the need to optimize feature selection for better generalization. On the other hand, when the number of features was reduced to three or four, the performance dropped dramatically, with accuracy falling to 70.37% and 71.39%, respectively. This indicates that these subsets did not have enough information to reliably classify.

Table 6. Evaluation of optimized heart disease detection using PSO-based feature selection: Dataset 2. Wilcoxon signed-rank test p-values compare each feature selection configuration against the baseline (all features).

Number of features	Accuracy	Precision	Recall	F1 score	p-value
All features	0.7397	0.7608	0.7005	0.7294	-
7	0.7395	0.7572	0.7064	0.7309	0.875
3	0.7037	0.7324	0.6436	0.6851	0.125
4	0.7139	0.7512	0.6410	0.6918	0.125
9	0.7411	0.7613	0.7037	0.7313	0.125
8	0.7406	0.7640	0.6975	0.7292	0.875

By comparing both datasets, PSO-based feature selection consistently improved performance by

identifying the most relevant features, reducing computational complexity while maintaining or even enhancing predictive accuracy. The results highlight that removing irrelevant or redundant features enhances generalization and prevents overfitting, making feature selection a crucial step in medical applications where interpretability and efficiency are essential. Future work could explore the integration of PSO with deep learning architectures to further refine the importance of characteristics and optimize hybrid models for heart disease detection.

Analyzing the statistics for Dataset 1 shows that feature selection affected model performance differently. The feature subsets with six, eight, and nine features had Wilcoxon p-values of 0.125 and 0.250, indicating that they performed similarly to the baseline model trained on all features. The model trained with nine features performed closest to the baseline (p-value = 0.125), showing that a well-optimized feature selection technique can reduce dimensionality without compromising the model's efficacy. However, feature sets with fewer features (for example, seven features, p = 0.250) had slightly higher variability, suggesting that feature selection procedures could be improved. The Wilcoxon test of Dataset 2 showed moderate changes from the baseline for feature subsets of three features (p = 0.125), four features (p = 0.125), and nine features (p = 0.125). Using "seven features" (p = 0.875) and "eight features" (p = 0.875) did not differ significantly from using all features. This shows that these two feature selections preserved the predictive capacity of the model while lowering the input complexity. These results show that a well-chosen subset of features can perform as well as the whole feature set, improving computational economy without losing accuracy.

4.5. Comprehensive evaluation: Federated learning with optimized feature selection and classifier integration

To take full advantage of the benefits of federated learning (FL), we used an FL-based strategy that used the best-performing classifier and the optimized feature set that we gained from previous tests. Federated learning improves privacy and security by training models on different devices without sharing data centrally. This is especially beneficial in healthcare applications, where data sensitivity is an important concern. We demonstrated the usefulness of our proposed model by comparing its results with those of other state-of-the-art techniques to identify heart disease in both data sets.

The findings, shown in Tables 7, show that, in terms of classification accuracy, our federated fearning-based model, refined by feature selection, exceeded earlier approaches. Our proposed model outperformed the ensemble model of Nazari et al. [42] based on GA (88.43%) and Mokeddem et al. [43] GA + Naive Bayes technique (85.50%). With an accuracy of 92.25%, this notable progress emphasizes how well feature selection, improved classifier selection, and federated learning together provide a stronger and more accurate model for detecting heart disease. Similarly, in Dataset 2, our federated learning-based model achieved an accuracy of 73.58%, slightly outperforming the GA-ANN method proposed by Arroyo et al. [44] (73.43%). Although the difference in performance is marginal in this case, it reinforces the viability of federated learning in maintaining accuracy while addressing privacy concerns in healthcare data.

The results show that, centralized machine learning approaches, integrating federated learning with optimal feature selection and classifier selection frequently increases models' performance. Thus, it is interesting to note that there were several cases in which the results obtained by centralized learning were only somewhat better than those obtained by federated learning. Given the nature of FL, where data are scattered among numerous clients and results in smaller training datasets for every node, this

discrepancy is expected, even if it is not very great.

Table 7. Comprehensive evaluation: Federated learning with optimized feature selection and classifier integration.

Dataset 1		
	Methods	Accuracy (%)
Proposed model	Federated Based on feature selection	92.25
Nazari et al. [42]	Ensemble model based on GA	88.43
Mokeddem et al. [43]	NB+GA	85.50
Dataset 2		
	Methods	Accuracy (%)
Proposed model	federated Based on feature selection	73.58
Arroyo et al. [44]	ANN+GA	73.43

Particularly in the medical field, where data exchange is highly sensitive and limited, federated learning has more advantages than any performance gap, even if one exists. Training models jointly without disclosing patient data guarantees compliance with data privacy criteria and builds confidence in medical applications using artificial intelligence. Furthermore, as our dataset is not very large, distributing it among several customers in federated learning reduces the available training data to each client, therefore affecting learning efficiency. However, the findings show that FL may continue to achieve competitive accuracy while simultaneously guaranteeing privacy protection, proving that it is still promising.

Moreover, our method outperforms the genetic algorithm (GA)-based techniques, suggesting that using PSO for feature selection combined with a better classifier produces a more exact feature subset that increases predictive power. Although the improvement in Dataset 2 is less obvious, it shows the adaptability of FL, demonstrating its competitiveness even in comparison with deep learning-based solutions such as GA-ANN.

Finally, by eliminating centralized data collection, FL improves privacy but increases the training time, computational cost, and communication efficiency. FL involves many local training iterations and communication with a central aggregator, unlike standard machine learning methods that train models on a single dataset. This increases client training time and requires careful aggregation frequency tuning to reduce the network overhead.

Although less accurate than centralized models, FL-based models perform competitively and improve data privacy, according to our study. The FL communication overhead depends on the model update size and communication rounds. Our strategy avoids superfluous parameter exchanges to improve performance and ensure only significant updates to the global model.

FL also faces data heterogeneity across clients, where local datasets affect model updates. This can slow convergence, requiring more aggregation rounds to stabilize learning. Our results show that FL-based training preserves privacy without degrading performance. The modest accuracy trade-off is offset by FL's improved security and feasibility in real-world healthcare settings, where regulatory constraints restrict data exchange.

FL adds computing overhead and connectivity constraints, but its ability to train models without revealing sensitive patient data makes it vital for medical AI applications. Compression methods can improve the communication efficiency, and federated deep learning can improve models' performance in heterogeneous medical datasets.

4.6. Impact of dataset bias on model generalization

The impact of data set bias on generalization is an important factor to consider, despite the fact that our models show competitive precision. In Dataset 2, there is a small imbalance between the classes, which could result in a slight bias in the predictions. This bias would require additional validation of the balanced data sets. In a similar vein, the large gender imbalance that exists between the two datasets shows that models may perform better on male patients than they do on female patients. This presents a potential concern in clinical applications where it is crucial to achieve equivalent prediction performance across demographics.

Reweighting and oversampling are two examples of bias mitigation approaches that should be investigated in future research to solve these constraints. These techniques will ensure that models learn equally from all groups of patients. Transfer learning and domain adaptation techniques could also help adapt models that were trained on a single data set to other populations with different distributions. Alternately, fairness-aware training is a viable approach that incorporates adversarial debiasing strategies to minimize the discrepancies that exist between individual demographic groups.

The need to ensure that prediction models are both fair and strong cannot be overstated in light of the growing integration of AI in medical diagnostics. Evaluation of model fairness across a variety of demographic groupings, validation of performance on datasets that are geographically diverse, and investigation of methods to eliminate data representation bias in training should be the primary focus of next-generation research. These procedures will ensure that artificial intelligence is ethically and reliably deployed in the healthcare industry, thus increasing the applicability of predictive models to a wider range of populations.

4.7. Computational cost and deployment feasibility

There are a number of computational issues that arise when federated learning (FL) is implemented in real-world healthcare settings. These challenges include the amount of time required for training and inference, the scalability limits, and the availability of resources. We performed an analysis of the computational cost of our suggested method as well as practical deployment issues to determine whether it is feasible.

4.7.1. Scalability considerations

The scalability of FL is affected by network latency, device availability, and the amount of computational power available at local nodes. However, FL offers benefits that protect patients' privacy by maintaining the decentralization of patient data sources. FL must be able to accept a wide range of client device capabilities in real-world healthcare settings. These capabilities can range from high-performance hospital servers to low-power edge devices such as mobile health monitors. Adaptive aggregation techniques, which reduce the frequency of model updates based on local performance improvements, can be used to alleviate the higher communication overhead associated

with FL. Additionally, federated dropout techniques might be developed to allow clients with lower resources to train only a subset of the model. This would reduce the amount of computational strain that these clients have to bear while still contributing to the global model.

4.7.2. Model deployment challenges and mitigation strategies

When using FL-based heart disease detection systems in real-world applications, there are several issues that need to be solved, including. limitations on bandwidth and concerns regarding security and privacy. Because FL requires frequent contact between clients and the central aggregator, it results in increased bandwidth utilization. Compression strategies, such as quantized updates and sparsified gradients, can be utilized to achieve optimal performance in this regard.

Despite the fact that FL improves data privacy, it is still susceptible to assaults such as data poisoning and model inversion. Secure aggregation methods, such as homomorphic encryption and differential privacy, can be used to prevent the leakage of sensitive data. There are many hospitals and clinics that rely on traditional AI systems, which require centralized data processing. Integration with existing medical systems is something that needs to be done. The deployment of FL requires modifications of the infrastructure to facilitate decentralized learning without interrupting the workflows already in place. In spite of these obstacles, the results of our experiments show that FL achieves an accuracy comparable with that of centralized models while also providing considerable advantages in terms of privacy. Optimization of FL for real-time medical applications, reducing computational overhead, and improving the model's robustness in diverse healthcare contexts will be the main focus of future developments.

5. Conclusions

In this study, we explored various machine learning approaches to enhance the accuracy and efficiency of heart disease detection. Our method integrates classical classifiers, ensemble learning, feature selection, and FL to improve predictive performance while addressing crucial data privacy and distributed learning challenges.

We examined different classifiers, including random forest, LightGBM, SVM, KNN, AdaBoost, and deep learning models such as CNN and LSTM. Our findings indicate that ensemble-based methods, particularly random forest and LightGBM, consistently outperformed individual classifiers, demonstrating their effectiveness in improving robustness and generalization. Incorporating voting (hard or soft) and stacking further reinforced predictive performance, highlighting the benefits of ensemble strategies in medical diagnostics. Using PSO for feature selection, we improved the efficiency of the model by ensuring that only the most relevant features were utilized. Choosing the most effective subset of predictive features, this optimization process reduced the number of redundant activities and improved the classification performance. Furthermore, we implemented FL to keep our predictive power intact while improving personal information protection. Our technology, built on FL, effectively enabled decentralized learning, ensuring that sensitive medical data remained on local devices while contributing to advances in global models. However, the FL technique is a realistic alternative for real-world healthcare applications because of its privacy protection and scalability advantages. This is despite presenting a few problems, such as heterogeneity in data distribution and a limited amount of local training data.

In the future, research should focus on extending FL models with deep learning architectures such as LSTM or federated CNNs to continue improving the accuracy of feature extraction and classification. It is also possible that the efficiency of the model could be improved by optimizing feature selection strategies using hybrid metaheuristic approaches. An inquiry could also be directed in a helpful direction by addressing the heterogeneity of data in FL through the implementation of specific aggregation algorithms. In conclusion, implementing and validating our FL-based heart disease detection system in clinical settings would provide more information on its practical application, hence strengthening its potential as a solution that is both secure and mindful of privacy in the context of AI-driven medical diagnostics.

Data availability

The data used in this study are openly accessible at the following link: `https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset\`
`https://www.kaggle.com/datasets/mexwell/heart-disease-dataset/data`

Author contributions

All authors of this article have contributed equally. All authors have read and approved the final version of the manuscript for publication

Use of Generative AI tools declaration

The authors declare that they have not used AI tools in the creation of this article.

Acknowledgments

This work was funded by the Deanship of Graduate Studies and Scientific Research at Jouf University under grant No. (DGSSR-2024-02-01182).

Conflict of interest

All authors declare that there are no competing interests.

References

1. S. Mendis, I. Graham, J. Narula, Addressing the global burden of cardiovascular diseases; need for scalable and sustainable frameworks, *Glob. Heart*, **17** (2022), 48. https://doi.org/10.5334/gh.1139

2. A. Ala, A. Goli, Incorporating machine learning and optimization techniques for assigning patients to operating rooms by considering fairness policies, *Eng. Appl. Artif. Intell.*, **136** (2024), 108980. https://doi.org/10.1016/j.engappai.2024.108980

3. N. Chaithra, B. Madhu, Classification models on cardiovascular disease prediction using data mining techniques, *J. Cardiovas. Dis. Diagn.*, **6** (2018), 10000348. https://doi.org/10.4172/2329-9517.1000348

4. K. W. Johnson, J. T. Soto, B. S. Glicksberg, K. Shameer, R. Miotto, M. Ali, et al., Artificial intelligence in cardiology, *JACC*, **71** (2018), 2668–2679.

5. S. Kodati, R. Vivekanandam, Analysis of heart disease using in data mining tools orange and weka, *Glob. J. Comput. Sci. Technol.*, **18** (2018), 16–22.

6. K. S. Shalet, V. Sabarinathan, V. Sugumaran, V. J. S. Kumar, Diagnosis of heart disease using decision tree and svm classifier, *Int. J. Appl. Eng. Res.*, **10** (2015), 598–602.

7. K. Uyar, A. ˙Ilhan, Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks, *Procedia Comput. Sci.*, **120** (2017), 588–593. https://doi.org/10.1016/j.procs.2017.11.283

8. S. Khader Basha, D. Roja, S. Santhj Priya, L. Dalavi, S. Srinivas Vellela, V. Reddy, Coronary heart disease prediction and classification using hybrid machine learning algorithms, In: *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, IEEE, 2023. https://doi.org/10.1109/ICIDCA56705.2023.10099579

9. R. Jahed, O. Asser, A. Al-Mousa, Using personal key indicators and machine learning-based classifiers for the prediction of heart disease, In: *2023 International Conference on Smart Computing and Application (ICSCA)*, IEEE, 2023. https://doi.org/10.1109/ICSCA57840.2023.10087430

10. A. Rajdhan, A. Agarwal, M. Sai, D. Ravi, P. Ghuli, Heart disease prediction using machine learning, *Int. J. Res. Technol.*, **9** (2020), 659–662.

11. C. Das, C. Das, M. Hossain, A. Rahman, H. Hossen, R. Hasan, Heart disease detection using ml, In: *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, IEEE, 2023. https://doi.org/10.1109/CCWC57344.2023.10099294

12. K. Pytlak, Indicators of heart disease (2022 update), 2022. Available from: `https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease`.

13. S. Chopra, N. Karla, R. Rani, Identification of cardiovascular disease using machine learning and ensemble learning, In: *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, IEEE, 2023. https://doi.org/10.1109/ICIDCA56705.2023.10099508

14. Z. Li, D. Zhou, L. Wan, J. Li, W. Mou, Heartbeat classification using deep residual convolutional neural network from 2-lead electrocardiogram, *J. Electrocardiol.*, **58** (2020), 105–112. https://doi.org/10.1016/j.jelectrocard.2019.11.046

15. L. B. Marinho, N. de MM Nascimento, J. W. M. Souza, M. V. Gurgel, P. P. R. Filho, V. H. C. de Albuquerque, A novel electrocardiogram feature extraction approach for cardiac arrhythmia classification, *Future Gener. Comput. Syst.*, **97** (2019), 564–577. https://doi.org/10.1016/j.future.2019.03.025

16. S. Pandya, T. Gadekallu, P. Reddy, W. Wang, M. Alazab, Infusedheart: A novel knowledge-infused learning framework for diagnosis of cardiovascular events, *IEEE Trans. Comput. Soc. Syst.*, **9** (2022), 1778–1788. https://doi.org/10.1109/TCSS.2022.3151643

17. S. Mohan, C. Thirumalai, G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques, *IEEE Access*, **7** (2019), 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

18. R. Arun, N. Deepa, Heart disease prediction system using naive bayes, *Int. J. Pure Appl. Math.*, **119** (2018), 3053–3065.

19. K. Prasanna, N. Challa, J. Nagaraju, Heart disease prediction using reinforcement learning technique, In: *2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, IEEE, 2023. https://doi.org/10.1109/ICAECT57570.2023.10118232

20. S. Bagavathy, V. Gomathy, S. S. Rani, M. Murugesan, K. Sujatha, M. K. Bhuvana, Early heart disease detection using data mining techniques with hadoop map reduce, *Int. J. Pure Appl. Math.*, **119** (2018), 1915–1920.

21. A. Abdellatif, H. Abdellatef, J. Kanesan, C. O. Chow, J. H. Chuah, H. M. Gheni, An effective heart disease detection and severity level classification model using machine learning and hyperparameter optimization methods, *IEEE Access*, **10** (2022), 79974–79985. https://doi.org/10.1109/ACCESS.2022.3191669

22. H. Ramesh, R. Pathinarupothi, Performance analysis of machine learning algorithms to predict cardiovascular disease, In: *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, IEEE, 2023. https://doi.org/10.1109/I2CT57861.2023.10126428

23. M. Shail, R. Sreeja, S. Zainab, P. S. Sowmya, T. Akshay, S. Sindhu, Improving accuracy of heart disease prediction through machine learning algorithms, In: *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, IEEE, 2023. https://doi.org/10.1109/ICIDCA56705.2023.10100244

24. L. Ali, A. Niamat, J. Khan, N. A. Golilarz, X. Xiong, A. Noor, et al., An optimized stacked support vector machines based expert system for the effective prediction of heart failure, *IEEE Access*, **7** (2019), 54007–54014. https://doi.org/10.1109/ACCESS.2019.2909969

25. T. Mahmud, A. Barua, M. Begum, E. Chakma, S. Das, N. Sharmen, An improved framework for reliable cardiovascular disease prediction using hybrid ensemble learning, In: *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, IEEE, 2023. https://doi.org/10.1109/ECCE57851.2023.10101564

26. B. Tama, S. Im, S. Lee, Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble, *Biomed Res. Int.*, **2020** (2020), 9816142. https://doi.org/10.1155/2020/9816142

27. O. Yildirim, A novel wavelet sequence based on a deep bidirectional lstm network model for ecg signal classification, *Comput. Biol. Med.*, **96** (2018), 189–202. https://doi.org/10.1016/j.compbiomed.2018.03.016

28. N. Fitriyani, M. Syafrudin, G. Alfian, J. Rhee, Hdpm: An effective heart disease prediction model for a clinical decision support system, *IEEE Access*, **8** (2020), 133034–133050. https://doi.org/10.1109/ACCESS.2020.3010511

29. D. Shah, S. Patel, S. K. Bharti, Heart disease prediction using machine learning techniques, *SN Comput. Sci.*, **1** (2020), 345. https://doi.org/10.1007/s42979-020-00365-y

30. A. Jain, K. Kumar, R. Tiwari, N. Jain, V. Gautam, N. K. Trivedi, Machine learning-based detection of cardiovascular disease using classification and feature selection, In: *2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT)*, IEEE, 2023. https://doi.org/10.1109/CSNT57126.2023.10134672

31. A. Shukla, I. Khan, V. Sharma, M. Soni, S. Gupta, A. Kumar, A novel prediction system to diagnose heart disease, In: *2023 International Conference on Inventive Computation Technologies (ICICT)*, IEEE, 2023. https://doi.org/10.1109/ICICT57646.2023.10133988

32. G. Bathla, P. Singh, R. Singh, E. Cambria, R. Tiwari, Intelligent fake reviews detection based on aspect extraction and analysis using deep learning, *Neural Comput. Appl.*, **34** (2022), 20213–20229. https://doi.org/10.1007/s00521-022-07531-8

33. G. Varshini, A. Ramya, C. Sravya, V. Kumar, B. K. Shukla, Improving heart disease prediction of classifiers with data transformation using pca and relief feature selection, In: *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*, IEEE, 2023. https://doi.org/10.1109/ICEARS56392.2023.10085401

34. G. Saranya, A. Pravin, A comprehensive study on disease risk predictions in machine learning, *Int. J. Elect. Comput. Eng.*, **10** (2020), 4217–4225. https://doi.org/10.11591/ijece.v10i4.pp4217-4225

35. M. Singh, L. M. Martins, P. Joanis, V. K. Mago, Building a cardiovascular disease predictive model using structural equation model & fuzzy cognitive map, In: *International Conference on Fuzzy Systems (FUZZ)*, IEEE, 2016, 1377–1382. https://doi.org/10.1109/FUZZ-IEEE.2016.7737850

36. K. Sen, B. Verma, Heart disease prediction using a soft voting ensemble of gradient boosting models, randomforest, and gaussian naive bayes, In: *2023 4th International Conference for Emerging Technology (INCET)*, IEEE, 2023. https://doi.org/10.1109/INCET57972.2023.10170399

37. K. Gola, S. Arya, Satin bowerbird optimization-based classification model for heart disease prediction using deep learning in e-healthcare, In: *2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW)*, IEEE, 2023. https://doi.org/10.1109/CCGridW59191.2023.00063

38. J. Patel, T. Upadhyay, S. Patel, Heart disease prediction using machine learning and data mining technique, *IJCSC*, **7** (2015), 129–137.

39. A. Pandita, S. Yadav, S. Vashisht, A. Tyagi, Review paper on prediction of heart disease using machine learning algorithms, *Int. J. Res. Appl. Sci. Eng. Technol.*, **9** (2021).

40. A. Akella, S. Akella, Machine learning algorithms for predicting coronary artery disease: efforts toward an open-source solution, *Future Sci. OA*, **7** (2021), FSO698.

41. A. Ala, V. Simic, D. Pamucar, N. Bacanin, Enhancing patient information performance in internet of things-based smart healthcare system: Hybrid artificial intelligence and optimization approaches, *Eng. Appl. Artif. Intell.*, **131** (2024), 107889. https://doi.org/10.1016/j.engappai.2024.107889

42. M. Nazari, H. Emami, R. Rabiei, A. Hosseini, S. Rahmatizadeh, Detection of cardiovascular diseases using data mining approaches: Application of an ensemble-based model, *Cogn. Comput.*, **16** (2024), 2264—2278. https://doi.org/10.1007/s12559-024-10306-z

43. S. Mokeddem, B. Atmani, M. Mokaddem, Supervised feature selection for diagnosis of coronary artery disease based on genetic algorithm, preprint paper, 2013. https://doi.org/10.48550/arXiv.1305.6046

44. J. C. T. Arroyo, A. J. P. Delima, An optimized neural network using genetic algorithm for cardiovascular disease prediction, *J. Adv. Inf. Technol.*, **13** (2022), 95–99. https://doi.org/10.12720/jait.13.1.95-99

AIMS Press

© 2025 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)