

Heart disease approach using modified random forest and particle swarm optimization

Khalidou Abdoulaye Barry¹, Youness Manzali¹, Rachid Flouchi², Mohamed Elfar¹

¹LPAIS Laboratory, Faculty of Sciences, Université Sidi Mohamed Ben Abdellah, Fez, Morocco

²Laboratory of Microbial Biotechnology and Bioactive Molecules, Faculty of Science and Technologies, Université Sidi Mohamed Ben Abdellah, Fez, Morocco

Article Info

Article history:

Received Mar 7, 2024

Revised Nov 1, 2024

Accepted Nov 14, 2024

Keywords:

Feature selection

Heart disease

Machine learning

Particle swarm optimization

Random forest

ABSTRACT

For the past two decades, heart disease has been classified as one of the main causes of mortality globally. Fortunately, most researchers focused on data mining techniques, which play an important role in accurately predicting heart disease to develop their models. In this paper, by combining particle swarm optimization (PSO) and modified random forest (MRF), a new approach (PSO-MRF) is proposed to predict heart disease. The main purpose is to select the important features after the bootstrap method for each decision tree in the random forest, and then optimize the MRF by the PSO algorithm. The experiments are carried out using the publicly accessible UCI heart disease datasets. Thorough experimental analysis demonstrates that our approach has outperformed the random forest algorithm as well as many other classifiers. This model helps doctors and researchers improve the diagnosis and treatment of heart disease, resulting in more prompt, accurate patient care.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Khalidou Abdoulaye Barry

LPAIS Laboratory, Faculty of Sciences, Université Sidi Mohamed Ben Abdellah

Fez 30000, Morocco

Email: khalidou.barry@usmba.ac.ma

1. INTRODUCTION

Heart diseases are the leading cause of death worldwide, and their incidence continues to rise [1]. The main causes of heart disease are narrowing or blockages in the coronary arteries, which provide blood to the heart [2]. The World Health Organization reports that close to 23.6 million people die from cardiovascular illnesses, with coronary heart disease and cerebral stroke accounting for 82 percent of these fatalities [3]. It has been estimated that, more than 30 million individuals would die from heart disease by 2040. Moreover, according to the American Heart Association, approximately 121.5 million adults in the United States suffer from cardiovascular disease [4]. Thus, it is critical to apply data mining and machine learning approaches to predict the probability of developing heart disease to help physicians and clinicians, who face several challenges in precisely and rapidly detecting heart disease. Besides, many researchers have conducted experiments to help doctors make decisions for heart disease prediction using machine learning techniques such as random forest [5], support vector machine, and K nearest neighbor [6]. However, when these classic algorithms are employed alone, they typically perform worse than when they are adjusted, either by improving the algorithm in question using meta-heuristic algorithms such as particles warm optimization (PSO) [7] and genetic algorithm (GA) [8] or by adding an additional step in the algorithm process that you would like to improve, for example as in our case, the random forest algorithm. Consequently, this work aims to employ the PSO algorithm to optimize

the modified random forest (MRF) approach, which will be discussed later. The paper's main contributions include determining the model's most effective hyperparameters, such as the number of estimators, maximum depth, and minimum number of samples required to divide a node. In addition, feature selection based on correlation was used in the random forest method. Finally, numerous metrics are employed to evaluate the proposed approach.

The rest of this study is organized as follows: section 2 outlines the related works. Section 3 offers the methodology of the proposed PSO-MRF model for heart disease prediction. Section 4 presents experimental results along with a discussion. Section 5 provides a conclusion and proposals for further work.

2. RELATED WORKS

This section reviews several machine learning algorithms for predicting heart disease. Recently, Asadi *et al.* [9] introduced a novel and extremely effective cardiac disease prediction model based on multi-objective PSO and random forest, utilizing a diverse dataset. However, the classification accuracy of their proposed approach needs to be further enhanced. In the same context, Latha and Jeeva [10] enhanced the accuracy of heart disease risk prediction using several classifiers. Similarly, a novel method has been proposed for heart disease classification using rough sets and fuzzy rule-based classification with an adaptive genetic algorithm, the experiments were done with the publicly available University of California Irvine (UCI) heart disease datasets [8]. Furthermore, three attribute selection approaches were used to extract the best set of features from the Cleveland heart dataset for different machine-learning models [11].

Moreover, Kadhim and Radhi [12] presented an early detection technique for heart disease that employs machine learning techniques such as random forest, support vector machines, K-nearest neighbor, and decision tree. Also, according to Henni *et al.* [13], two approaches for coronary artery disease prediction are proposed: the first optimizes a random forest model by hyperparameter adjustment. The second approach employs case-based reasoning (CBR) methodology. According to Ozcan and Peker [14], the classification and regression tree (CART) algorithm was used to detect cardiac disease and generate decision rules for clarifying correlations between input and output information. Additionally, a quantum machine learning-based ensemble learning method for heart disease prediction was developed [15]. The authors reported that quantum-enhanced machine learning algorithms outperform traditional techniques for diagnosing heart failure. In addition, the best first search was used to determine the most significant features, followed by seven machine learning algorithms for modeling, among these seven, the random forest gave the best accuracy, which is 90% [16].

Similarly, an effective machine learning-based detection system for identifying heart disease was created employing six machine-learning algorithms [17]. The experiments were conducted using datasets from the Cleveland and IEEE Data ports. The soft voting ensemble classifier approach was used for all six models on both datasets, resulting in accuracies of 93.44% and 95%, respectively. Moreover, two classifiers for heart disease prediction such as the random forest classifier, and the XG Boost classifier, were developed and enhanced using hyperparameter optimization techniques such as grid search, randomized search, and genetic programming (TPOT classifier) [18]. Besides, K-nearest neighbor and random forest are utilized to classify individuals as having heart disease or not [19]. The prediction accuracy of K-nearest neighbor is 86.885%, whereas the random forest technique is 81.967%. In the same context, Spencer *et al.* goal in [20] was to find a combination of filter and classification approaches that perform well together to improve heart disease prediction. The authors reported that using Chi-squared feature selection with the Bayes Net classifier resulted in the best accurate model (85.0% accuracy, 84.73% precision, and 85.56% recall). As you will have understood, most of the aforementioned articles, along with articles [21]–[24], emphasize the importance of feature selection to improve model performance.

Hence, according to this literature, we noticed a weakness performance in these studies as well and one of the most commonly used approaches is random forest, which has several advantages. Nevertheless, this approach can be further improved by using meta-heuristic algorithms or revising its procedure. In this work, we selected the features most correlated with the target variable for each decision tree of the random forest and then we used the PSO method to determine the optimal max-depth, the min samples split, and the n-estimators for the MRF.

3. METHOD

3.1. Data collection and pre-processing

For the present study, we worked with the cleveland dataset from the UCI machine learning repository. The original dataset contains 76 raw variables and 303 rows, however, only thirteen features and one attribute as an output class are commonly utilized to predict cardiovascular disease [25]. The original class value was a multi-class variable with a value range of 0–4. The 0 value indicates the absence of heart disease, whereas values 1–4 indicate the existence of heart disease and its stage. We converted the class value from a multi-class variable to a binary-class variable, as done in [26]. As shown by Table 1, the details of the patients show that they are aged between 29 and 77 in the data range column, along with other relevant information. Furthermore, after the data collection step, we did the preprocessing data which is the most crucial step before applying the proposed model since it would eliminate missing values and duplicates, allowing the different machine-learning algorithms to produce more accurate predictions. In this instance, we used a standard scaler since it is the most effective approach employed in the suggested study to normalize the data into common scales.

Table 1. Feature description

Number	Feature name	Description	Data range
1.	age	age in years	[29,77]
2.	sex	Gender	0=female, 1=male
3.	cp	Chest pain type	1=typical angina, 2=atypical angina, 3=non-anginal pain, 4=asymptomatic
4.	trestbps	The blood pressure levels during rest	[94,200]
5.	Chol	Serum cholesterol in mg/dl	[126,564]
6.	fbs	The blood sugar levels while Fasting	0=false, 1=true
7.	restecg	Resting electrocardiographic results	0=normal, 1=having ST-T abnormality, 2=showing probable or definite left ventricular hypertrophy
8.	thalach	Maximum heart Rate attained	[71,202]
9.	exang	Angina caused by physical activity	0=no, 1=yes
10.	oldPeak	ST depression resulting from exercise relative to rest	[0,6.2]
11.	slope	The slope of the peak exercise ST segment	1=up-sloping, 2=flat, 3=down-sloping
12.	ca	Number of main vessels (0-3) colored with flourosopy	0-3
13.	thal	Defect type	3=normal, 6=fixed defect, 7=reversable defect
14.	target	target(Heart disease)	0=absence, 1=presence

3.2. Proposed model

The proposed approach uses the PSO algorithm and MRF to identify and diagnose cardiovascular disease risk on a real dataset. The following subsection will offer more information on random forests. The PSO approach is a well-known population-based metaheuristics technique for solving optimization issues [27]. It is inspired by social activities including swarming, bird flocking, and fish schooling. The PSO algorithm works on the concept that each answer is depicted as a vector known as a particle. Every single particle is allocated a place in the field of search to investigate the best solutions. Consequently, each particle has a velocity. During each movement, each particle's velocity and location are updated depending on its own experiences and those of its neighbors. The particle's best prior position is recorded as the personal best (P_{bst}), and the best position acquired by the population thus far is referred to as (G_{bst}). Based on P_{bst} and G_{bst} , PSO looks for optimal solutions by updating each particle's velocity and location. In (1) updates the velocity of particle k in the swarm at the $(i + 1)^{th}$ iteration, while in (2) updates the position of each particle k at each iteration ($i + 1$).

$$V_k(i + 1) = V_k(i) + C1rd1(P_{bst,i}^k - X_k(i)) + C2rd2(G_{bst,i} - X_k(i)) \quad (1)$$

$$X_k(i + 1) = X_k(i) + V_k(i + 1) \quad (2)$$

$$V_k(i + 1) = wV_k(i) + C1rd1(P_{bst,i}^k - X_k(i)) + C2rd2(G_{bst,i} - X_k(i)) \quad (3)$$

Where, $V_k(i + 1)$ stands for particle k's velocity at the $(i + 1)^{th}$ iteration, $X_k(i)$ represents the particle's position, w is the inertia weight employed to balance the local and global exploitation, c1 and c2 are the real acceleration coefficients that control how much the global and individual best positions should influence the particle's velocity, and rd1 and rd2 are uniformly distributed random numbers in the range 0 and 1, used to

preserve a suitable level of randomness. Moreover, multiple decision trees are constructed in a random forest classification utilizing diverse random subsets of the data and attributes. In the present work, we selected the correlated feature with the target variable for each random subset to generate the corresponding decision tree, and then the majority-voting approach was utilized to provide a unique output. The proposed model has been optimized using the PSO technique; as exemplified by Figure 1 which consist of two parts: the first is MRF, where each tree is accompanied by arrows simulating their processes. The second part is the PSO algorithm, which optimizes the MRF.

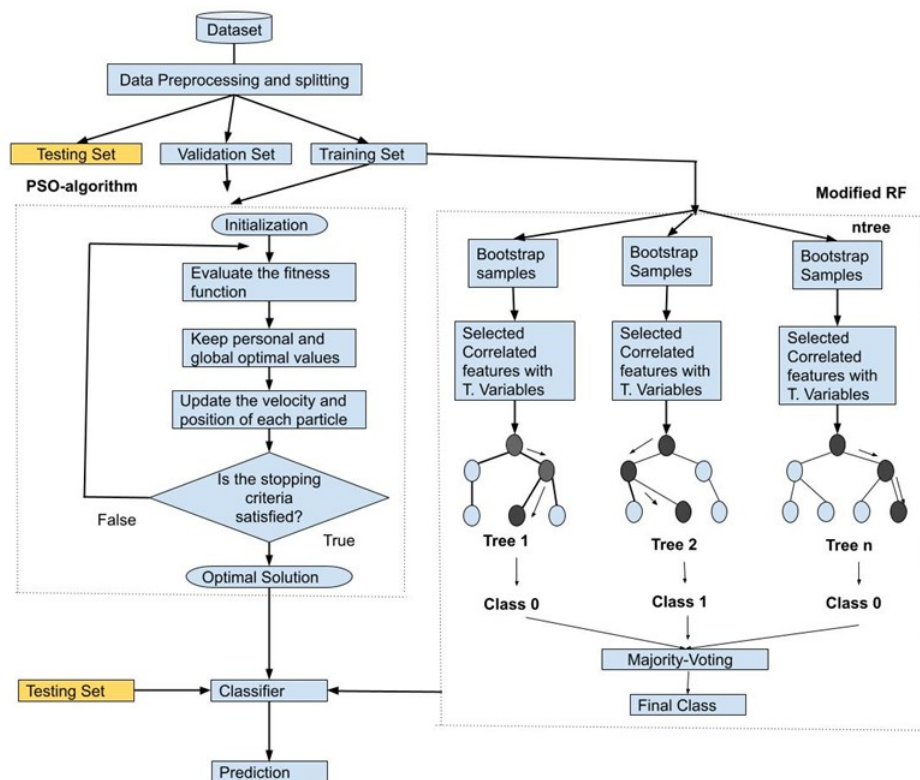


Figure 1. The proposed model architecture

3.3. Comparison models

3.3.1. Decision tree

It is a tree structure created to solve classification or regression problems. In the tree structure, a branching node evaluates a characteristic, while the leaf node provides information on an individual class label [28]. To assess performance using decision tree, we should adopt the following steps [2]: compute the Gini index; split the dataset and analyze each split, then choose the best split; create a decision tree.

3.3.2. Random forest

Random forest models use ensemble learning to make predictions based on the most frequent results from several model runs. The user defines the number of independent models, or decision trees, within the forest. The random forest method relies on a random selection of input data to create distinct trees, which is known as the bootstrap technique [29]. To get a general idea of the structure of this technique, see Figure 2, and some details were given in [30].

3.3.3. Support vector machines

Support vector machine is a machine learning technique that optimizes the space between positive and negative data points closest to the decision hyperplane in an N-dimensional space, especially when there

are multiple classes to identify [31]. This method is typically used for both linear and non-linear data separation [32]. And it recognizes points in one dimension, lines in two, planes in three, and hyperplanes in four dimensions [28]. Moreover, support vector machine is mathematically expressed as follows:

$$If Y_j = 1; cx_j + b \geq 1 \quad (4)$$

$$If Y_j = -1; cx_j + b \leq -1 \quad (5)$$

$$For all j; Y_j(cx_j + b) \geq 1 \quad (6)$$

In the formula, (x) represents a vector point, whereas (c) is both a weight and a vector. As a result, the data in (4) must always be greater than zero, yet the data in (5) must always be less than zero.

3.3.4. Naïve Bayes

The naïve Bayes algorithm is a wonderful technique that improves classification accuracy with statistical and probabilistic techniques, making it applicable to diverse datasets. Naïve Bayes may address several issues such as classification, clustering, association, prediction, and estimate [33]. The naïve Bayes theorem relies on the Bayes formula:

$$P(lb/ft) = \frac{P(lb)P(ft/lb)}{P(ft)} \quad (7)$$

Where ft: features; lb: labels; $P(lb/ft)$ is the posterior probability of class; $P(ft/lb)$ is the predictor's likelihood given the class probability; $P(lb)$ denotes the prior probability of class; $P(ft)$ denotes the predictor's prior probability.

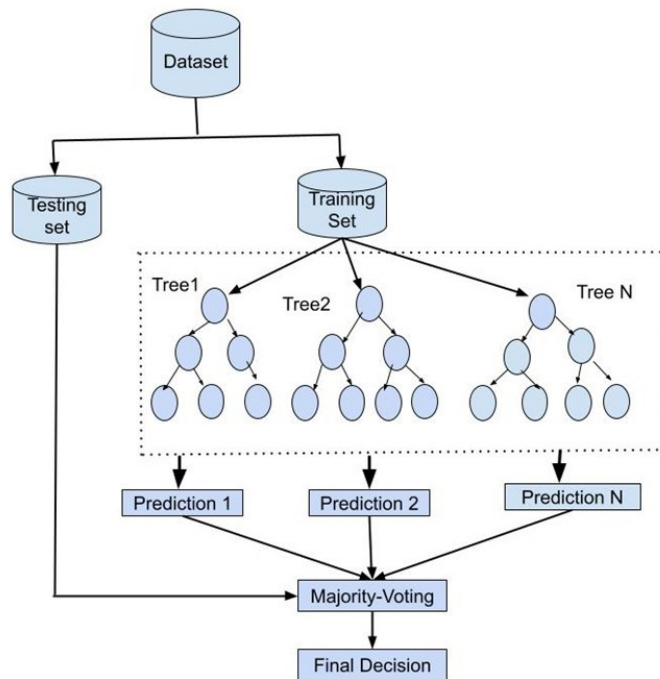


Figure 2. The random forest architecture

3.4. Experimental setup

This study uses the appropriate performance evaluation metrics for classification problems, including accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUC-ROC), and geometric-mean (G-mean), to assess the proposed model's performance. The confusion matrix includes four classification performance indices. The definitions for these are given below: T_{posi} refers to true positive

(the number of subjects correctly classified as positive); T_{nega} refers to true negative (the number of subjects correctly classified as negative); F_{posi} refers to false positive (the number of subjects incorrectly classified as positive); F_{nega} refers to false negative (the number of subjects incorrectly classified as negative).

Besides, the previously mentioned evaluation measures are calculated as follows:

- Accuracy indicates the percentage of correctly identified samples in the dataset. The mathematical definition as in (8):

$$Accuracy = \frac{T_{posi} + T_{nega}}{T_{posi} + T_{nega} + F_{posi} + F_{nega}} \quad (8)$$

- Precision indicates how many individuals with heart disease were identified by the model as having the disease, given by (9):

$$Prec = \frac{T_{posi}}{T_{posi} + F_{posi}} \quad (9)$$

- Sensitivity/recall was used to calculate the ratio of correctly classified positive samples to total positive occurrences, which gives (10):

$$Sensitivity(Sensi)/recall = \frac{T_{posi}}{T_{posi} + F_{nega}} \quad (10)$$

- The F1 score determines the harmonic mean of the model's precision and sensitivity (or recall) scores, given by (11):

$$F1 - score = \frac{2Prec \times Sensi}{Prec + Sensi} \quad (11)$$

- AUC-ROC is another evaluation metric that offers an overall measure of the performance of a classification model, given by (12):

$$AUC - ROC = \frac{1}{2} \left(\frac{T_{posi}}{T_{posi} + F_{nega}} + \frac{T_{nega}}{T_{nega} + F_{posi}} \right) \quad (12)$$

- G-mean is a metric that uses two additional metrics, such as sensitivity and specificity, to get a single score, given by (13):

$$G - mean = \sqrt{(Sensitivity \times Specificity)} \quad (13)$$

4. RESULTS AND DISCUSSION

We assessed the PSO-MRF classifier's performance in predicting heart disease using the cleveland dataset from the UCI repository. First, after preprocessing the data, we divided it into training and testing samples in an 8:2 proportion. We divided the training set in an 8:2 ratio into training data and validation data. Training data is used to train the MRF model, which is then combined with validation data to train the PSO algorithm, as seen in the flowchart, and testing data is used to evaluate performance. The proposed PSO-MRF model has been evaluated based on accuracy, precision, recall, f1-score, AUC, and G-mean. Their scores are summarized in Table 2.

Table 2. Performance metrics of PSO-MRF model

Performance metrics	Score (%)
Accuracy	88.52
Precision	86.11
Sensitivity/Recall	91.18
F1-score	90.14
AUC-ROC	93.40
G-mean	87.57

According to Table 2, we can notice that the PSO-MRF model obtained an accuracy of 88.52%, demonstrating that it is capable of identifying patients with and without heart disease. The 86.11% precision means that 86.11% of the patients predicted to be positive for heart disease were positive. The 91.18% recall rate shows that the model accurately identified 91.18% of all true positive cases of heart disease. The F1-score, which balances precision and sensitivity or recall, is 90.14%. Furthermore, the ROC-AUC value is 0.9340, indicating that the PSO-MRF model is highly discriminative. The G-mean score, which accounts for both sensitivity and specificity, is 0.8757. We evaluated the PSO-MRF classifier's performance against multiple machine learning models often used for heart disease prediction, including standard random forest, decision trees, naive Bayes, and support vector machines. As follows, a comparison of these four models with the proposed model is made using the first three evaluation metrics given in Figure 3, the precision-recall curve comparison in Figure 4, and the ROC-curve in Figure 5. All comparisons demonstrated that the PSO-MRF model outperformed the other approaches used in this work. Accordingly, we summarize the positive findings obtained based on researching the optimum of different random forest parameters such as max-depth, min samples split, and n-estimators. As an instance, in our case, the max-depth was assigned 7, the min samples split was 3, and the n-estimators were 143. As a result, our study found that the PSO-MRF outperformed a classic random forest algorithm in predicting heart disease. In the end, this work was compared with previous studies as shown in Table 3.

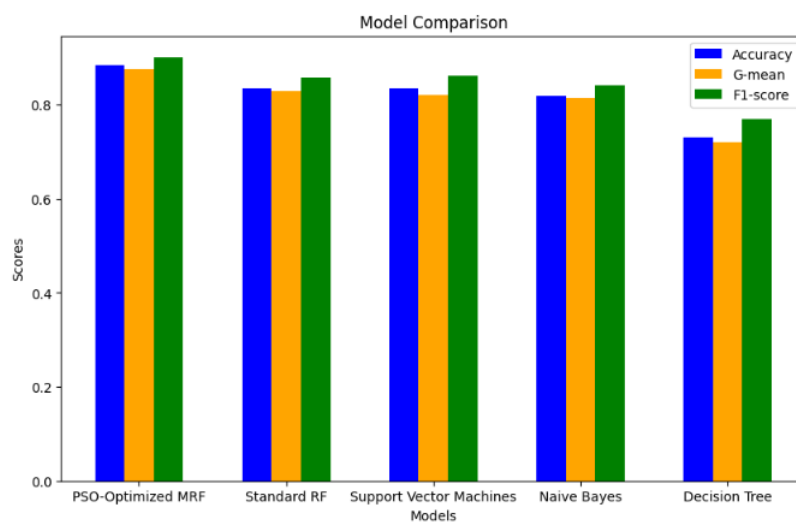


Figure 3. The models comparison

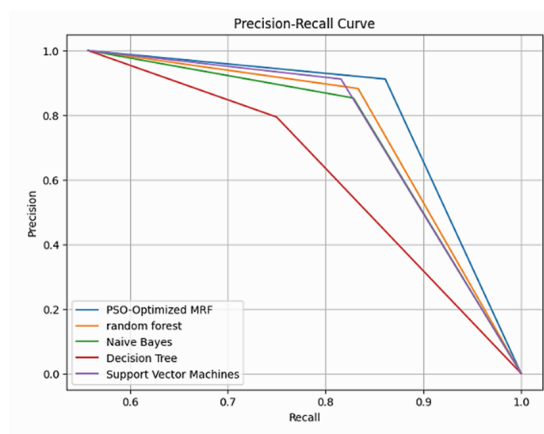


Figure 4. The precision-recall curve comparison

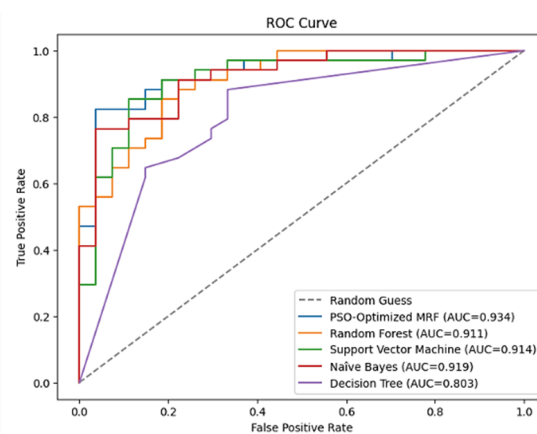


Figure 5. The ROC curve comparison

Table 3. Comparison of this work results with previous studies in terms of accuracy

Study	Data	Method	Accuracy (%)
Enriko <i>et al.</i> [34]	The Cleveland heart dataset	Weighted k-nearest neighbors	81.90
Rubini <i>et al.</i> [35]	The Cleveland dataset	random forest, naive Bayes, and logistic regression	84.81
Mohan <i>et al.</i> [36]	The Cleveland dataset	The hybrid random forest with a linear model	88.40
Djerioui <i>et al.</i> [37]	The Cleveland dataset	component analysis and support vector machine	85.43
Latha and Jeeva [10]	The Cleveland dataset	Ensemble classification	85.48
Dwivedi [38]	The Statlog dataset	Logistic regression	85
Amin <i>et al.</i> [39]	The Statlog dataset	Naive Bayes and logistic regression	87.41
Saqlain <i>et al.</i> [40]	The Statlog dataset	Mean Fisher score-based feature selection algorithm (MFSFSA) and support vector machine	81.19
Mary and Sebastian [41]	The Cleveland heart and two others data were used	Naive Bayes classifier, random forest, and random tree	86.8132
Nugroho <i>et al.</i> [42]	The Cleveland dataset	Support vector machines	88
Proposed model	The Cleveland dataset	The MRF and PSO algorithm	88.52

5. CONCLUSION

Heart disease is a dangerous disease, by the way, millions of people all around the world are suffering from heart disease. Hence, an effective heart disease prediction model can play a crucial role in saving patients' lives. This paper presents a novel technique to improving the random forest method's performance. Indeed, firstly, we focused on determining the features correlated with the target variable from different subsets generated by the bootstrap method for each decision tree in the random forest rather than using a feature with a weak or no contribution to the decision tree in question. Secondly, we optimized the MRF using the PSO method to find the optimal max-depth, min-sample split, and n-estimators. Furthermore, the results reveal that the PSO-MRF model outperformed the four standalone classifiers (support vector machine, naïve Bayes, decision tree, and standard random forest), as well as many other previous studies. When comparing classic random forest to the proposed method, it is clear that identifying correlated features with the target variable and selecting appropriate max-depth, min-sample split, and n-estimators significantly impact random forest performance. Hopefully, this method would make the physician's task simpler. In the future, this study can be considerably enhanced by evaluating various meta-heuristic algorithms such as the whale optimization method, the Antlion algorithm, and the adaptive bee colony algorithm. Additionally, this model can be evaluated on a variety of additional medical datasets.

REFERENCES




- [1] M. M. Ahsan and Z. Siddique, "Machine learning-based heart disease diagnosis: A systematic literature review," *Artificial Intelligence in Medicine*, vol. 128, pp. 102-289, 2022, doi: 10.1016/j.artmed.2022.102289.
- [2] S. I. Ayon, M. M. Islam, and M. R. Hossain, "Coronary artery heart disease prediction: a comparative study of computational intelligence techniques," *IETE Journal of Research*, vol. 68, no. 4, pp. 2488-2507, 2022, doi: 10.1080/03772063.2020.1713916.
- [3] T. R. Ramesh, U. K. Lilhore, M. Poongodi, S. Simaiya, A. Kaur, and M. Hamdi, "Predictive analysis of heart diseases with machine learning approaches," *Malaysian Journal of Computer Science*, no. 1, pp. 132-138, 2022.
- [4] E. J. Benjamin *et al.*, "Heart disease and stroke statistics - 2018 update: a report from the American Heart Association," *Circulation*, vol. 137, no. 12, pp. E67-E492, 2018, doi: 10.1161/CIR.0000000000000558.
- [5] M. Pal and S. Parija, "Prediction of heart diseases using random forest," *Journal of Physics: Conference Series*, vol. 1817, no. 1, 2021, doi: 10.1088/1742-6596/1817/1/012009.
- [6] D. A. Anggoro, "Comparison of accuracy level of support vector machine (SVM) and k-nearest neighbors (KNN) algorithms in predicting heart disease," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 5, pp. 1689-1694, 2020, doi: 10.30534/ijeter/2020/32852020.
- [7] T. M. Shami, A. A. El-Saleh, M. Alswaiti, Q. Al-Tashi, M. A. Summakieh, and S. Mirjalili, "Particle swarm optimization: a comprehensive survey," *IEEE Access*, vol. 10, pp. 10031-10061, 2022, doi: 10.1109/ACCESS.2022.3142859.
- [8] G. T. Reddy, M. P. K. Reddy, K. Lakshmana, D. S. Rajput, R. Kaluri, and G. Srivastava, "Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis," *Evolutionary Intelligence*, vol. 13, no. 2, pp. 185-196, 2020, doi: 10.1007/s12065-019-00327-1.
- [9] S. Asadi, S. E. Roshan, and M. W. Kattan, "Random forest swarm optimization-based for heart diseases diagnosis," *Journal of Biomedical Informatics*, vol. 115, 2021, doi: 10.1016/j.jbi.2021.103690.
- [10] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, 2019, doi: 10.1016/j.imu.2019.100203.
- [11] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand, "Heart disease risk prediction using machine learning classifiers with attribute evaluators," *Applied Sciences*, vol. 11, no. 18, 2021, doi: 10.3390/app11188352.
- [12] M. A. Kadhim and A. M. Radhi, "Heart disease classification using optimized machine learning algorithms," *Iraqi Journal for Computer Science and Mathematics*, vol. 4, no. 2, pp. 31-42, 2023, doi: 10.52866/ijcs.2023.02.02.004.

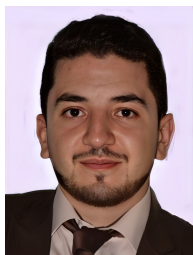
- [13] F. Henni, B. Atmani, F. Atmani, and F. Saadi, "Improving coronary artery disease prediction: use of random forest, feature importance and case-based reasoning," *International Journal of Decision Support System Technology*, vol. 15, no. 1, 2023, doi: 10.4018/ijdsst.319307.
- [14] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," *Healthcare Analytics*, vol. 3, 2023, doi: 10.1016/j.health.2022.100130.
- [15] G. Abdulsalam, S. Meshoul, and H. Shaiba, "Explainable heart disease prediction using ensemble-quantum machine learning approach," *Intelligent Automation and Soft Computing*, vol. 36, no. 1, pp. 761–779, 2023, doi: 10.32604/iasc.2023.032262.
- [16] M. I. Hossain et al., "Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison," *Iran Journal of Computer Science*, vol. 6, no. 4, pp. 397–417, 2023, doi: 10.1007/s42044-023-00148-7.
- [17] N. Chandrasekhar and S. Peddakrishna, "Enhancing heart disease prediction accuracy through machine learning techniques and optimization," *Processes*, vol. 11, no. 4, 2023, doi: 10.3390/pr11041210.
- [18] R. Valarmathi and T. Sheela, "Heart disease prediction using hyper parameter optimization (HPO) tuning," *Biomedical Signal Processing and Control*, vol. 70, 2021, doi: 10.1016/j.bspc.2021.103033.
- [19] A. Garg, B. Sharma, and R. Khan, "Heart disease prediction using machine learning techniques," *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, 2021, doi: 10.1088/1757-899X/1022/1/012046.
- [20] R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, "Exploring feature selection and classification methods for predicting heart disease," *Digital Health*, vol. 6, 2020, doi: 10.1177/2055207620914777.
- [21] A. K. Gárate-Escamila, A. H. El Hassani, and E. Andrés, "Classification models for heart disease prediction using feature selection and PCA," *Informatics in Medicine Unlocked*, vol. 19, 2020, doi: 10.1016/j.imu.2020.100330.
- [22] N. R. Kolukula, P. N. Pothineni, V. M. K. Chinta, V. G. Boppana, R. P. Kalapala, and S. Duvvi, "Predictive analytics of heart disease presence with feature importance based on machine learning algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 2, pp. 1070–1077, 2023, doi: 10.11591/ijeecs.v32.i2.pp1070-1077.
- [23] A. Ullah, S. A. Khan, T. Alam, S. Luma-Osmami, and M. Sadie, "Heart disease classification using various heuristic algorithms," *International Journal of Advances in Applied Sciences*, vol. 11, no. 2, pp. 158–167, 2022, doi: 10.11591/ijaas.v11.i2.pp158-167.
- [24] K. Dissanayake, and M. G. M. Johar, "Comparative study on heart disease prediction using feature selection techniques on classification algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2021, no. 1, 2021, doi: 10.1155/2021/5581806.
- [25] R. R. Sarra, A. M. Dinar, and M. A. Mohammed, "Enhanced accuracy for heart disease prediction using artificial neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 1, pp. 375–383, 2023, doi: 10.11591/ijeecs.v29.i1.pp375-383.
- [26] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPm: an effective heart disease prediction model for a clinical decision support system," *IEEE Access*, vol. 8, pp. 133034–133050, 2020, doi: 10.1109/ACCESS.2020.3010511.
- [27] M. Jain, V. Saihjal, N. Singh, and S. B. Singh, "An overview of variants and advancements of pso algorithm," *Applied Sciences*, vol. 12, no. 17, 2022, doi: 10.3390/app12178392.
- [28] C. Gupta, A. Saha, N. V. S. Reddy, and U. D. Acharya, "Cardiac disease prediction using supervised machine learning techniques," *Journal of Physics: Conference Series*, vol. 2161, no. 1, 2022, doi: 10.1088/1742-6596/2161/1/012013.
- [29] P. Josso, A. Hall, C. Williams, T. L. Bas, P. Lusty, and B. Murton, "Application of random-forest machine learning algorithm for mineral predictive mapping of fe-mn crusts in the world ocean," *Ore Geology Reviews*, vol. 162, 2023, doi: 10.1016/j.oregeorev.2023.105671.
- [30] P. Ghosh et al., "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.
- [31] E. I. Elsedimy, S. M. M. AboHashish, and F. Algarni, "New cardiovascular disease prediction approach using support vector machine and quantum-behaved particle swarm optimization," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 23901–23928, 2024, doi: 10.1007/s11042-023-16194-z.
- [32] N. M. Ali, N. A. A. Aziz, and R. Besar, "Comparison of microarray breast cancer classification using support vector machine and logistic regression with lasso and boruta feature selection," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 20, no. 2, pp. 712–719, 2020, doi: 10.11591/ijeecs.v20.i2.pp712-719.
- [33] A. Saleh, N. Dharshinni, D. Perangin-Angin, F. Azmi, and M. I. Sarif, "Implementation of recommendation systems in determining learning strategies using the naïve bayes classifier algorithm," *Sinkron*, vol. 8, no. 1, pp. 256–267, 2023, doi: 10.33395/sinkron.v8i1.11954.
- [34] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart disease prediction system using k-nearest neighbor algorithm with simplified patient's health parameters," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 8, no. 12, pp. 59–65, 2016.
- [35] P. E. Rubini, C. A. Subasini, A. V. Katharine, V. Kumaresan, S. Gowdhankumar, and T. M. Nithya, "A cardiovascular disease prediction using machine learning algorithms," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 2, pp. 904–912, 2021.
- [36] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
- [37] M. Djerioui, Y. Brik, M. Ladjal, and B. Attallah, "Neighborhood component analysis and support vector machines for heart disease prediction," *Ingenierie des Systemes d'Information*, vol. 24, no. 6, pp. 591–595, 2019, doi: 10.18280/isi.240605.
- [38] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Computing and Applications*, vol. 29, no. 10, pp. 685–693, 2018, doi: 10.1007/s00521-016-2604-1.
- [39] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telematics and Informatics*, vol. 36, pp. 82–93, 2019, doi: 10.1016/j.tele.2018.11.007.
- [40] S. M. Saqlain et al., "Fisher score and matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines," *Knowledge and Information Systems*, vol. 58, no. 1, pp. 139–167, 2019, doi: 10.1007/s10115-018-1185-y.
- [41] T. R. S. Mary and S. Sebastian, "Predicting heart ailment in patients with varying number of features using data mining techniques," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 4, pp. 2675–2681, 2019, doi: 10.11591/ijece.v9i4.pp2675-2681.




- [42] K. S. Nugroho, A. Y. Sukmadewa, A. Vidiyanto, and W. F. Mahmudy, "Effective predictive modelling for coronary artery diseases using support vector machine," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, pp. 345–355, 2022, doi: 10.11591/ijai.v11.i1.pp345-355.

BIOGRAPHIES OF AUTHORS






Khalidou Abdoulaye Barry    was born in Djewol, Mauritania, in 1995. He received his bachelor's degree in mathematics and computer science from the Faculty of Science and Technology of Nouakchott in Mauritania. In addition, he received his master's degree in computer science, precisely systems intelligence, and decision (MSID) from the Faculty of Science at Dhar El Mehraz (Sidi Mohamed Ben Abdellah University, Fez, Morocco), in 2020. He is currently a Ph.D. student in the Laboratory of Applied Physics, Statistics, and Computer Science (LPAIS) at FSDM. His research areas of interest include artificial intelligence and digital signal processing. He can be contacted at email: khalidou.barry@usmba.ac.ma.






Youness Manzali    was born in Fez, Morocco, in 1991. He received his master's degree in information systems, networks, and multimedia (SIRM) from the Faculty of Sciences (FSDM), USMBA Fez, in 2015. He is currently a Ph.D. student in the LPAIS at FSDM. Youness is a proficient researcher whose significant contributions to the field are evidenced by the publication of scientific papers in internationally recognized peer-reviewed journals. His primary research focuses on machine learning algorithms and data analysis. He can be contacted at email: younes.manzali@usmba.ac.ma.



Pr. Rachid Flouchi    was born in Taza, Morocco, in 1983. He is a doctor in microbiology and biotechnology from the Faculty of Sciences and Techniques of Fez, Morocco. He is currently a lecturer at the Higher Institute of Nursing Professions and Health Technologies, Taza Annex, Fez. Author of several indexed international scientific articles and associate member of the Laboratory of Microbial Biotechnology and Bioactive Molecules at the Faculty of Sciences and Technologies of Fez. His research areas of interest include: antimicrobial activities, biotechnology, bioinformatics, and epidemiology. He can be contacted at email: rachid.flouchi@usmba.ac.ma.



Mohamed Elfar    is a Professor of Computer Science in the Faculty of Science at Dhar El Mehraz (Sidi Mohamed Ben Abdellah University, Fez, Morocco). He has authored numerous research papers, with his primary focus lying in artificial intelligence, databases, and big data analytics. He can be contacted at email: mohamed.elfar@usmba.ac.ma.