Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal **Vol. 5, No. 4, October 2023, pp: 250-260; eISSN: 2656-8632**

RESEARCH ARTICLE **OPEN ACCESS**

Manuscript received July 27, 2023; revised August 20, 2023; accepted September 21, 2023; date of publication October 30, 2023
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeemi.v5i4.322

How to cite: Muhammad Ridho Ansyari, Muhammad Itqan Mazdadi, Fatma Indriani, Dwi Kartini, Triando Hamonangan Saragih, "Implementation of Random Forest and Extreme Gradient Boosting in the Classification of Heart Disease Using Particle Swarm Optimization Feature Selection, vol. 5, no. 4, pp. 250-260, October 2023.

Implementation of Random Forest and Extreme Gradient Boosting in the Classification of Heart Disease Using Particle Swarm Optimization Feature Selection

Muhammad Ridho Ansyari, Muhammad Itqan Mazdadi, Fatma Indriani, Dwi Kartini, Triando Hamonangan Saragih

Computer Science Department, Lambung Mangkurat University, Banjarbaru, South Kalimantan, Indonesia

Corresponding Author: Muhammad Itqan Mazdadi (email: mazdadi@ulm.ac.id).

This work was supported by Lambung Mangkurat University for providing valuable resources and support.

ABSTRACT Heart disease is a condition that ranks as the primary cause of death worldwide. Based on available data, over 36 million people have succumbed to non-communicable diseases, and heart disease falls within the category of non-communicable diseases. This research employs a heart disease dataset from the UCI Repository, consisting of 303 instances and 14 categorical features. In this research, the data were analyzed using the classification methods XGBoost (Extreme Gradient Boosting) and Random Forest, which can be applied with PSO (Particle Swarm Optimization) as a feature selection technique to address the issue of irrelevant features. This issue can impact prediction performance on the heart disease dataset. From the results of the conducted research, the obtained values for the XGBoost (Extreme Gradient Boosting) model were 0.877, and for the Random Forest model, it was 0.874. On the other hand, in the model utilizing Particle Swarm Optimization (PSO), the obtained AUC values are 0.913 for XGBoost (Extreme Gradient Boosting) and 0.918 for Random Forest. These research results demonstrate that PSO (Particle Swarm Optimization) can enhance the AUC of heart disease prediction performance. Therefore, this research contributes to enhancing the precision and efficiency of heart disease patient data processing, which benefits heart disease diagnosis in terms of speed and accuracy.

INDEX TERMS XGBoost, Random Forest, PSO, Heart

I. INTRODUCTION

Heart disease is the main source of death overall every year and is a non-communicable disease. More than 17,9 million deaths occur every year worldwide because of heart disease, which can be preventable to reduce mortality rates due to heart disease [1]. However, medical diagnosis of heart disease requires the involvement of experts in the field. One effective approach to identifying and predicting heart disease is by harnessing machine learning algorithms [2]. Machine learning can overcome these limitations as a tool for conducting heart disease diagnosis. Machine learning possesses the capability to model by learning from data, akin to human learning, to decide if a patient is determined to have coronary illness or not [3].

The machine learning techniques that can be employed for classification are Random Forest and XGBoost (Extreme Gradient Boosting). Random Forest is one of the advancements of the Decision Tree method. Random Forest has the advantage of enhancing accuracy results and ameliorating Decision Tree methods that are prone to overfitting [4][5]. Random Forest is a supervised classification algorithm that leverages multiple classification trees. Classification is performed by passing each input vector down each tree randomly. The algorithm model is based on Decision

Trees, which is effective for cases involving categorical features in the data [6].

XGBoost is one of classification algorithms with decision trees as its base learner, and it is an evolution of the Gradient Tree Boosting algorithm based on ensemble algorithms. It efficiently addresses large-scale machine learning problems. XGBoost is versatile, capable of performing various tasks such as regression, classification, and ranking. It constitutes a tree ensembles algorithm comprising a collection of classification and regression trees (CART). XGBoost excels in terms of speed and memory utilization. Leveraging enhanced processor caching, multicore processing, and distributed parallel computing, the system operates faster than other popular algorithms [7]. However, despite the robustness of Random Forest and XGBoost, they can yield models that are overly complex or exhibit poor performance if not accompanied by proper feature selection. Therefore, PSO can be utilized as a feature selection method to enhance the classification performance.

PSO can address the issue of noise attribute in datasets, consequently enhancing the outcomes of classification [8]. In a research conducted by Batool, a PSO-based SVM classification method was applied in sensors technologies for human activity analysis. The research results demonstrated that the implementation of PSO in the SVM classification method led to an accuracy increase to 0.875[9]. Hence, the utilization of PSO feature selection is expected to enhance the classification performance outcomes in heart disease prediction.

In this research, the application of the XGBoost (Extreme Gradient Boosting) and Random Forest classification methods with Particle Swarm Optimization (PSO) as a feature selection technique is conducted to address the issue of irrelevant features (noise attributes). It is predicted that the Random Forest and XGBoost algorithms will categorize heart disease data more accurately and efficiently by using PSO, leading to increased precision. The point of this exploration is to research whether the execution of PSO can enhance the predictive performance of heart disease in the XGBoost and Random Forest classification methods using the heart disease dataset. The expected contributions of this research include: a. improving knowledge of how feature selection and classification techniques are applied in health datasets, especially in heart disease cases; b. helping medical professionals optimize decision-making through analysis; c. improving the accuracy of data evaluation through the use of the Random Forest algorithm and the XGBoost algorithm with Particle Swarm Optimization.

II. METHOD

The following is the research procedure that will be conducted. Figure 1 illustrates the flowchart of this research.

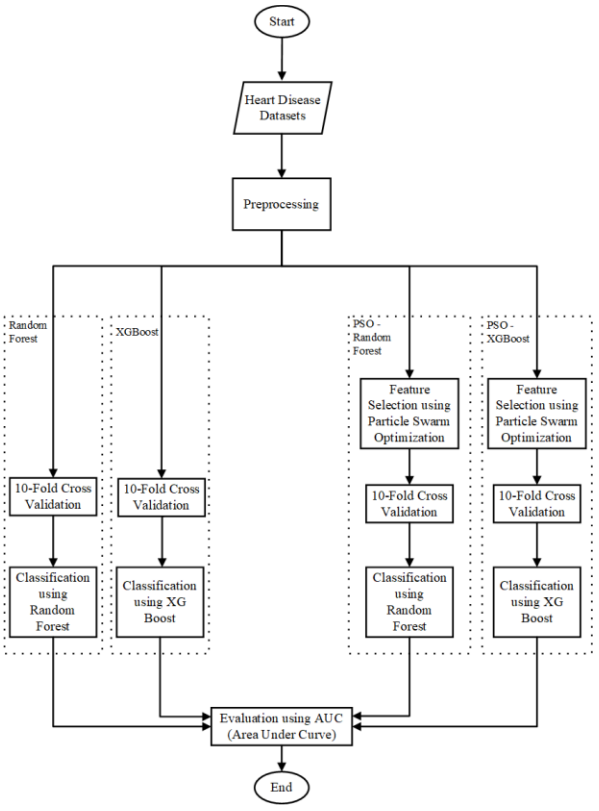


FIGURE 1. Research Flowchart

As shown in Figure 1, the workflow of this research starts with the assortment of the dataset, trailed by the phases of preprocessing and feature selection utilizing PSO. After highlight determination is applied to the dataset, the next step involves classification. In this research, there are two scenarios: first, classification using XGBoost (Extreme Gradient Boosting) and Random Forest without Particle Swarm Optimization; second, classification using XGBoost and Random Forest with Particle Swarm Optimization. The dataset is partitioned according to the 10-Fold Cross Validation rule and is evaluated using the AUC.

A. *DATA COLLECTION*

The research employed secondary data obtained from the UCI Repository website. This dataset can be seen at the following link https://archive.ics.uci.edu/dataset/45/heart+disease. This dataset pertains to heart disease issues and comprises a total of 303 instances. The heart disease dataset utilized consists of 14 attributes, with 13 attributes representing patient clinical statuses, used as predictive attributes, and the target class is the 14th attribute. These attributes are *Age, Sex, Cp, Trestbps, Chol, Fbs, Restecg, Thalac, Exang, Oldpeak, Slope, Ca, Thal, and Target.* The target attribute has only two values, 0 for non-cardiac condition and 1 for cardiac condition. According to the research [10] Table 1 presents a description of the research data attributes.

TABLE 1
Description of Research Data Attributes

No	Attribute	Description	Category
1	Age	Age in years	Numeric
2	Sex	Sex (01=male; 02=female)	Binary
3	Cp	Chest pain type	Nominal
4	Trestbps	Resting blood pressure	Numeric
5	Chol	Serum cholesterol in mg/dl	Numeric
6	Fbs	Fasting blood sugar> 120 mg/dl 01= true; 02=false	Binary
7	Restecg	Resting electrocardiographic results	Nominal
8	Thalach	Maximum heart rate achieved	Numeric
9	Exang	Exercise-induced angina	Binary
10	Oldpeak	ST depression induced by exercise relative to rest	Numeric
11	Slope	The slope of peak exercise ST segment	Nominal
12	Ca	Number of major vessels (0-3) colored by fluoroscopy	Nominal
13	Thal	3=normal; 6=fixed defect; 7=reversable defect	Nominal
14	Target	Diagnosis of heart disease	Binary

B. RANDOM FOREST

The Random Forest (RF) algorithm, as its name suggests, is a supervised classification algorithm that employs a random process to generate a forest. The number of trees within the forest directly affects the accuracy of the outcomes, with larger numbers of trees resulting in greater precision [11]. Random Forest classification is done by obtaining the majority class votes from the individual vote class trees [12]. One important benefit associated with RF relates to the fact that there is no need to prune individual trees, given the presence of multiple trees. However, the disadvantage is that due to the large number of trees, the ability to visualize them effectively is impaired [13]. This methodology is founded on two fundamental concepts: sampling by row and voting classification. After being resampled, the supplied records are submitted to the following base learner models for training. The idea of aggregating is a voting classifier where the test data output is chosen for the class that receives the most votes among the base learner models[14]. A generalized model for the Random Forest is depicted in Figure 2. Random Forest is one of the many Ensemble techniques created by Leo Breiman in 2001. It extends the Classification and Regression Tree (CART) method through incorporation of Bootstrap Aggregating (Bagging) and Random Feature Selection. The Random Forest method itself boasts several advantages, including yielding strong classification outcomes, minimizing error rates, and efficiently handling training data sets of substantial magnitude [15].

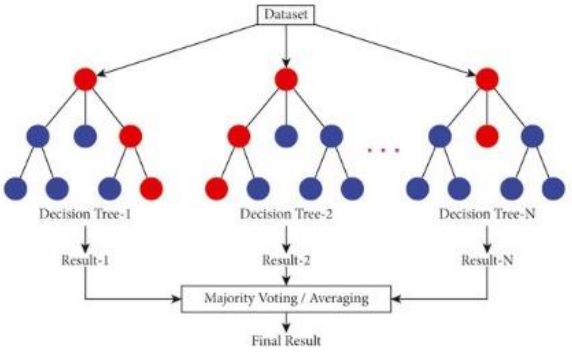


FIGURE 2. Random Forest Structure[14]

The Random Forest method produce an ensemble of random trees. The resulting class originates from the classification process, chosen from the most frequent class (mode) generated by the existing decision trees [16]. When making a collective classification decision, the presence of subpar trees can lead to accurate predictors generating erroneous predictions. The Random Forest operator generates an ensemble of random trees, and the resulting class in the classification process is determined by selecting the most frequent class (mode) generated across the existing random trees. To enhance the stability of importance measurements, it is recommend to utilize a substantial number of trees, particularly when research considers importance metrics and confronts a multitude of independent variables. Metrics such as Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG) are utilized within the context of Random Forest to quantify importance[17] The execution of Random Forest involves the following procedure:

1. Performing an irregular sampling of size n with replacement from dataset clusters, this step constitutes the bootstrap phase.
2. By employing a bootstrap example, trees are constructed until reaching their greatest size (without pruning). At every hub, the determination of a splitter is performed by haphazardly picking m logical indicators, where m << p, and afterward the best splitter is picked in view of these m indicators. This stage is alluded to as the irregular element choice stage.
3. Repeat stages 1 and 2 k times, bringing about the development of a backwoods comprising of k trees.

C. EXTREME GRADIENT BOOSTING (XGBoost)

XGBoost is a high level execution of the slope supporting calculation that utilizes decision trees as the basis for classification. It is widely employed due to its speed, efficacy, and scalability in solving various problems related to regression and classification[7]. The basic idea of this calculation is to iteratively change the learning boundaries to limit the expense capability. XGBoost (Extreme Gradient Boosting) employs a more disciplined model to construct tree structures, enabling it to achieve superior performance and reducing model complexity to prevent overfitting [18] [19].

XGBoost (Extreme Gradient Boosting) is a tree learning algorithm capable of handling sparse data or missing values. XGBoost offers advantages regarding speed and memory usage, such as optimized processor store use and backing for multicore handling, making the system run more than ten times faster than commonly used popular solutions. By analyzing each variable in XGBoost learning, it is possible to construct an accurate and efficient XGBoost model (Extreme Gradient Boosting) with or without the use of feature selection techniques such as PSO.

The Xgboost method employs the shrinkage technique [7] to group feeble learners and reduce the likelihood of model overfitting. This ensemble has the configuration shown below.

$$F_m(X) = F_{m-1}(X) + nf_m(X), 0 < n < 1 \quad (1)$$

Where f_m (X) is the mth iteration for creating the weak learner and f_m (X) is the mth iteration for creating the integrated learner. Since the parameter n has a significant negative correlation with the quantity of iterations, the generalization properties of the model are frequently improved when n has a value that is smaller.

D. PARTICLE SWARM OPTIMIZATION (PSO)

James Kennedy and Russell Eberhart developed the particle swarm optimization algorithm in 1995 as a stochastic optimization algorithm. The social behavior of fish, birds, insects, and other animal communities served as the model for the algorithm. An iterative technique called particle swarm looks through a population of particles for the best possible answer. In pursuit of food, these particles circle an area, with the particle nearest to the food indicating to the others to go in that direction. The orbiting particle closest to the target source will communicate with the others in order to approach it if any of them are closer than the initial particle. This process is repeated until one of the particles locates the sustenance (optimal solution)[20].

PSO is a developmental calculation with global optimization enlivened by the way of behaving of bird rushes and fish schools. Each bird is depicted as a particle addressing an answer for an issue and has a position (x) and velocity (v). Advantages of this algorithm include rapid convergence to the global best point, uncomplicated execution, a limited amount of adjustable parameters, and increased computational efficiency[21]. The Particle Swarm Optimization (PSO) algorithm consists of several sequential steps. Firstly, initialization randomly selects particles as search agents (x) and their velocities (v). Next, we apply the cost function to the particles to find the optimal global (Gbest) and local (Pbest) solutions. It is said that the best local location is the one with the lowest cost per particle. Furthermore, the optimal global solution is the location that, when all local solutions are added together, has the lowest cost. Finally, to update the particles, the following equation is applied:

$$v_{n+1} = v_n + c_1 r_1 (p_{best} - x_n) + c_2 r_2 (g_{best} - x_n) \quad (2)$$

$$x_{n+1} = x_n + v_{n+1} \quad (3)$$

Where c1 and c2 are constants, r1 and r2 are random integers, and n denotes the number of repetitions[22]. In equations 4 and 5, the following formula can be used to calculate the position displacement and particle velocity:

$$v_i(t) = v_i(t-1) + c_1 r_1 [Xp_{best}\, i - X_i(t)] + c_2 r_2 [xg_{best}\, i - xi\,(t)] \quad (4)$$

$$x_i\,(t) = x_i\,(t-1) + v_i(t) \quad (5)$$

In equations 4 and 5, These are the variables that are utilized to calculate the particle's displacement and velocity. In order to determine the particle's speed during the search for the best solution, Vit(t) stands for the velocity of particle i at iteration t. The solution attained by particle i at iteration t is represented by Xi(t), which is particle i's position at that time. Because of this, c1 and c2 are learning rate variables that show how much the change in particle velocity is influenced by social (group) and individual particle (cognitive) features. The particle's level of success in finding a solution is defined by c1, while the effect of its group members on the particle's search for a better solution is indicated by c2. Then, random values evenly distributed between intervals of 0 and 1 make up r1 and r2. By exploring the search space randomly, particles are able to identify more optimal solutions thanks to the stochastic component of these random numbers in the solution search process.

In the pursuit of the optimal solution, XPbest i indicates the most effective position of particle i to date. This is used to guide the movement of particles toward a more ideal solution and takes each particle's prior accomplishments into account. Meanwhile, XGbest i represents the best global position a group particle has achieved. This denotes the optimal solution that the group as a whole arrived at and gives particles direction to arrive at more optimal solutions jointly. In the PSO technique, it is possible to determine the displacement and velocity of particles by including each of these variables in equations 4 and 5. This is carried out repeatedly until the target outcome or the iteration limit is reached, which yields the best possible solution to the issue.

E. AREA UNDER THE ROC (RECEIVER OPERATING CHARACTERISTIC) CURVE

The ROC curve, frequently referred to AUC value, is a visual tool widely employed by researchers to evaluate prediction outcomes and compare two classification models. Figure 3 illustrates that the ROC is a two-layered chart with bogus up-sides as the even hub and genuine up-sides as the upward hub. The diagonal line that divides the ROC space illustrates that the area over the askew line demonstrates great grouping, while the region beneath the corner to corner line implies unfortunate order. True random guessing lies along the diagonal line, ranging from the lower left to the upper right.

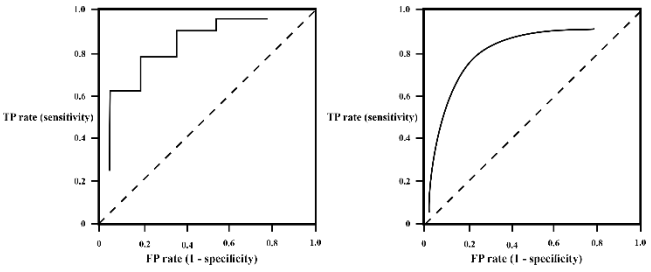


FIGURE 3. Example ROC Graph

AUC can be regarded as a probability calculated by the formula for the area under the curve. The categorization approach provides positive examples with higher scores than negative examples when one chooses positive and negative examples randomly. Consequently, AUC is a target for optimization because a higher number indicates a more effective classification technique [22]. The AUC value will always lie between 0 and 1, since the value ranges of both the x and y axes in square units range from 0 to 1. Random guessing would produce a corner-to-corner line with an area of 0.5 for values larger than 0.5, anywhere between 0 and 1. Table 2 is a list of the several groups that AUC values for data mining categorization can be split into[22] .

TABLE 2
Accuracy of classification results based on AUC values

AUC Values	Category
0.90-1.00	Excellent Classification
0.80-0.90	Good Classification
0.70-0.80	Fair Classification
0.60-070	Poor Classification
0.50-0.60	Failure

1. DATA COLLECTION

In this research, secondary data are employed. Secondary data is known as data acquired from third parties and not directly obtained from the research subjects. The data on heart disease used in this study can be accessed at the UCI Repository. With 13 attributes acting as predictor variables and a single attribute acting as the target variable, this data set has 303 total data points.

2. PREPROCESSING

In the following research, there are several missing data points. Therefore, the preprocessing step conducted involves removing missing values. This is in accordance with the journal [24] Generally, *missing values* do not significantly impact the entire dataset, especially if they represent a small percentage, say 1% of the total dataset. Hence, in the subsequent research, missing values are removed.

3. FEATURE SELECTION

As demonstrated in Figure 1, prior to entering the classification stage, the dataset undergoes a feature selection process. Highlight determination is led to address unessential elements that could be beneficial in enhancing the presentation of the characterization model. The component choice carried out in this research employs the PSO method. The particles within PSO traverse and search the solution space for the best solution [25]. With the selection of relevant features, it is anticipated that the performance of heart disease prediction can be enhanced. Therefore, feature weighting is carried out using PSO to obtain feature weight values going from 0 to 1. Figure 4 illustrates the RapidMiner scheme for utilizing the PSO (Particle Swarm Optimization) method.

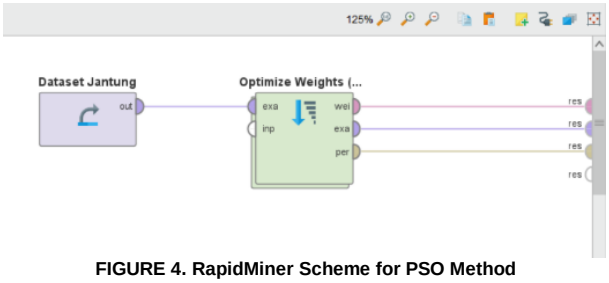


FIGURE 4. RapidMiner Scheme for PSO Method

The following is the process of feature weighting in the heart disease dataset, as depicted in Table 3.

TABLE 3
Feature weighting by PSO

Attribute	Weight
FBS	0
THALACH	0
CHOL	0.142
SLOPE	0.387
CA	0.707
RESTECG	0.828
SEX	0.903
CP	0.941
AGE	1
TRESTBPS	1
EXANG	1
OLDPEAK	1
THAL	1

Based on the feature selection process by Particle Swarm Optimization (PSO), features with a weight of 0 will be eliminated. In Table 3, there are 2 features with a weight of 0. Table 4 illustrates the features that have passed through the feature selection stage.

TABLE 4
Feature selection results by PSO

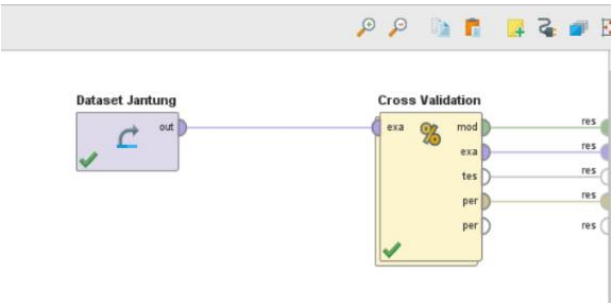
Attribute	Weight
CHOL	0.142
SLOPE	0.387
CA	0.707
RESTECG	0.828
SEX	0.903
CP	0.941
AGE	1
TRESTBPS	1
EXANG	1
OLDPEAK	1
THAL	1

4. CLASSIFICATION

The process of data analysis that involves determining a model or function to represent a concept or data class is referred to as classification. [26]. The objective of classification is to accurately predict categories within known data for each case. Classification algorithms can be applied to categorical data, while for numerical target data, predictive models utilized are regression algorithms [27]. In this stage, data is initially divided. The dataset is randomly partitioned into 10 subsets, with each subset containing nearly equal amounts of data and class proportions. *Cross-Validation* is employed as a performance evaluation technique to ensure the reliability of prediction outcomes [28]. Classification modeling is carried out using XGBoost and Random Forest.

a. XGBoost Classification

The classification was conducted on the heart disease dataset that had been separated into preparing and testing information. Subsequently, all training data was utilized to perform classification using the XGBoost (Extreme Gradient Boosting) model in RapidMiner Studio. Figure 5 illustrates the RapidMiner scheme for utilizing the XGBoost method.



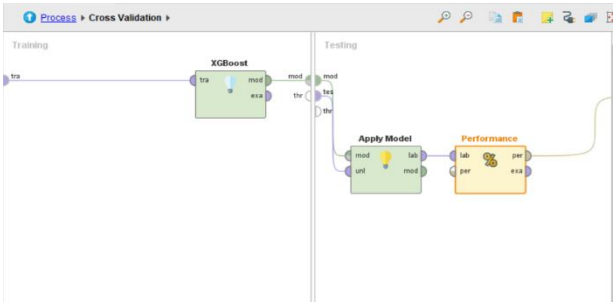


FIGURE 5. RapidMiner Scheme for XGBoost Method

b. XGBoost Classification with Particle Swarm Optimization (PSO)

In this phase, an experiment was conducted to initialize the number of particles, starting from the smallest particle count up to the point where the AUC value becomes relatively stable or decreases within a maximum of 10 iterations. Therefore, the particles for PSO were selected based on the highest attained AUC value. Table 5 presents the PSO particles on the heart disease dataset using the XGBoost (Extreme Gradient Boosting) model.

TABLE 5
PSO particles with Xgboost model

Dataset	Particle
Heart Disease	4

c. Random Forest Classification

The classification was conducted on the heart disease dataset that had been divided into training and testing data. Subsequently, all training data was employed to perform classification using the Random Forest model in RapidMiner Studio. Figure 6 illustrates the RapidMiner scheme for utilizing the Random Forest method.

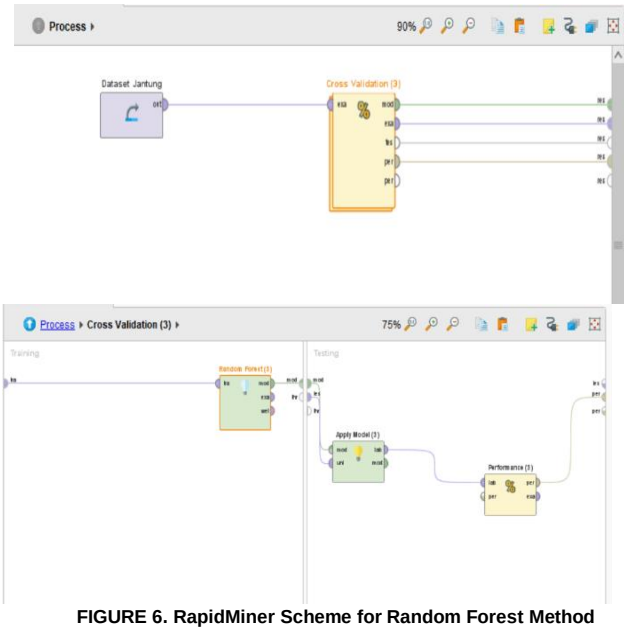


FIGURE 6. RapidMiner Scheme for Random Forest Method

d. Random Forest Classification with Particle Swarm Optimization (PSO)

In this phase, an experiment was conducted to initialize the number of particles, starting from the smallest particle count up to the point where the AUC value becomes relatively stable or decreases within a maximum of 10 iterations. Thus, the particles for PSO were selected based on the highest attained AUC value. Table 6 indicates the PSO particles on the heart disease dataset using the Random Forest model.

TABLE 6
PSO particles with random forest model

Dataset	Particle
Heart Disease	14

5. EVALUATION

The Area Under Curve (ROC) will be used to evaluate the models' effectiveness in predicting heart disease based on the heart disease dataset. AUC is chosen as the evaluation method because it is more suitable for assessing the performance value of predictions using both balanced and imbalanced datasets [29]. Table 7 illustrates the classification accuracy quality from the testing using AUC values.

TABLE 7
Accuracy of classification results based on AUC values

0.90 – 1.00	Excellent Classification
0.80 – 0.90	Good Classification
0.70 – 0.80	Fair Classification
0.60 – 0.70	Poor Classification
0.50 – 0.60	Failure

III. RESULTS

This research will present the evaluation results of AUC in heart disease prediction from the heart disease dataset. The models employed for predicting heart disease include XGBoost, XGBoost with 4 particles PSO (XGBoost + PSO4), Random Forest, and Random Forest with 14 particles PSO (RF + PSO14). The evaluation results comprise performance values indicated in Table 8.

TABLE 8
AUC result values for heart disease

Model	AUC Values
XGBoost	0.877
XGBoost + PSO 4	0.913
Random Forest	0.874
RF + PSO 14	0.918

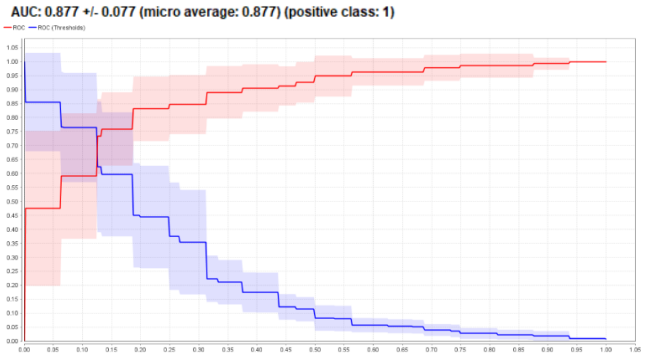


FIGURE 7. XGBoost ROC Curve

The ROC curve, as shown in Figure 7 above, was produced during the testing of the XGBoost (Extreme Gradient Boosting) approach. With an AUC value of 0.877, it was classified as a Good Classification because it fell within the range of 0.80-0.90.

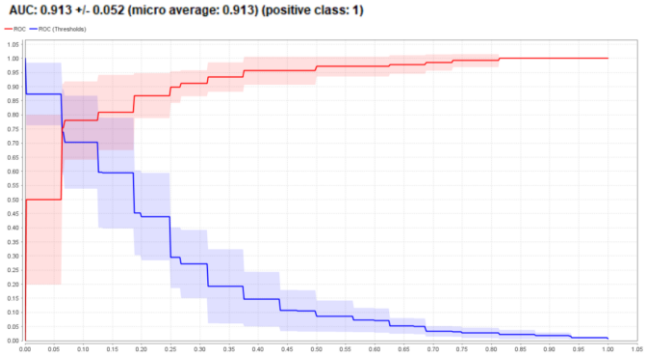


FIGURE 8. XGBoost ROC Curve with PSO (4 particles)

In the testing of the XGBoost method with PSO, using 4 particles, the ROC curve was obtained as observed in Figure 8 above, with an AUC value of 0.913, categorized as Excellent Classification due to falling within the range of 0.90-1.00.

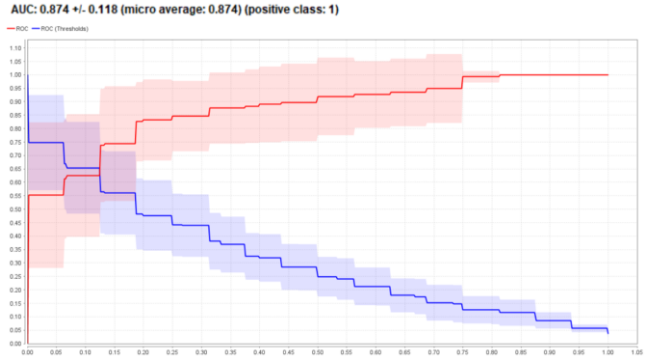


FIGURE 9. Random Forest ROC Curve

In the testing of the Random Forest method, the ROC curve was obtained as seen in Figure 9 above, with an AUC value of 0.874, categorized as Good Classification due to falling within the range of 0.80-0.90.

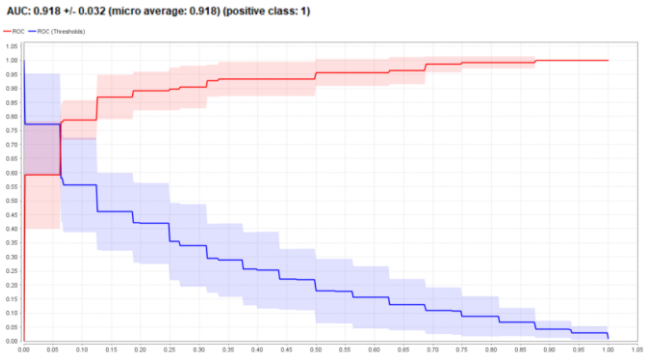


FIGURE 10. Random Forest ROC Curve with PSO (14 particles)

In the testing of the Random Forest method with Particle Swarm Optimization (PSO), using 14 particles, the ROC curve was obtained as depicted in Figure 10 above, with an AUC value of 0.918, categorized as Excellent Classification due to falling within the range of 0.90-1.00.

IV. DISCUSSION

This research employs a heart disease dataset for conducting heart disease prediction. The dataset is divided using a 10-fold cross-validation approach, wherein dataset is partitioned into 10 sets of data, with a distribution of 9 for training information and 1 for test information. The training information is utilized for the classification stage, while the testing data is employed to validate the model. This testing was conducted with 10 iterations. All subsets of data were alternately employed for both training and assessment purposes. Two classification techniques, XGBoost and Random Forest, were utilized in this study. In addition, feature selection using PSO was used to resolve noise attributes within the heart disease dataset, thereby improving the prediction of heart disease performance.

In PSO process, an experiment was conducted to initialize the number of particles in order to obtain the best-performing particles for both classifications, as indicated in Table 5 and Table 6. Subsequently, the models were evaluated using AUC to assess their performance in heart disease prediction. The obtained model performance was used to compare the baseline models of XGBoost (Extreme Gradient Boosting) and Random Forest with the models of XGBoost and Random Forest after integrating Particle Swarm Optimization (PSO).

In the baseline XGBoost (Extreme Gradient Boosting) model, the AUC evaluation results for heart disease prediction yielded an AUC value of 0.877. According to the general guidelines for AUC classification, the AUC evaluation results fall under the category of good classification.

In the baseline Random Forest model, the AUC evaluation results for heart disease prediction yielded an AUC value of 0.874. According to the general guidelines for AUC classification, the AUC evaluation results fall under the category of good classification.

In the XGBoost (and Random Forest models utilizing PSO, in the case of XGBoost with Particle Swarm Optimization, the AUC evaluation results for heart disease prediction yielded the best AUC value when employing 4 particles, with an AUC of 0.913. Meanwhile, in the Random Forest model with Particle Swarm Optimization, the best AUC value was obtained using 14 particles, yielding an AUC of 0.918. Based on the general guidelines for AUC classification, the AUC evaluation results fall under the category of excellent classification.

The heart disease dataset experiences an increase in AUC values when employing XGBoost with PSO and Random Forest with PSO, as compared to the baseline models of XGBoost and Random Forest. Table 9 illustrates the AUC evaluation results of the four models employed in heart disease prediction.

TABLE 9
AUC for Heart Disease Prediction

DATASET	XGB	RF	XGB + PSO	RF + PSO
Heart Disease	0.877	0.874	0.913	0.918

To assess which model exhibits the highest increase in AUC results from the heart disease dataset, a comparison of the AUC values among the four models can be conducted. Figure 11 illustrates the AUC comparison graph of the four classification models on the heart disease dataset. The results of the Random Forest and XGBoost tests improved in accordance with the feature selection and weighting using Particle Swarm Optimization. The Random Forest model with Particle Swarm Optimization exhibits the most superior AUC among the other models, as observed in Figure 11. The study's findings were compared with those of other investigations, and it was found that Random Forest and XGBoost, when combined with PSO feature selection, were more effective at classifying datasets pertaining to heart disease. Specifically, the combined approaches enhanced accuracy and AUC values compared to past research that utilized different classification algorithms or did not use PSO feature selection. The research results of Ajdani & Ghaffary, as well as Jiang, He, Ye & Zhang [30],[31], The accuracy and AUC values were not comparable despite the use of the same classification algorithms as this study. This study demonstrated that using Random Forest and XGBoost with PSO feature selection can enhance the classification performance of heart disease datasets by increasing their accuracy and area under the curve (AUC). This suggests that integrating methods with optimization improved the classification of survival rates for heart disease patients. In addition, this comparison aids in understanding the possibility for the techniques and algorithms utilized in this study to produce superior results compared to those of earlier research. However, the present study was limited by the use of a dataset with just a few of patients and features. This may have an effect on the generalizability of the results. In order to acquire more accurate and exhaustive results in future research, it is advised to use a larger and more diverse dataset that includes a greater number of patients and cardiac disease-related characteristics. This study effectively demonstrated that combining Random Forest with PSO and

XGBoost with PSO can improve classification performance in cardiac disease despite these limitations.

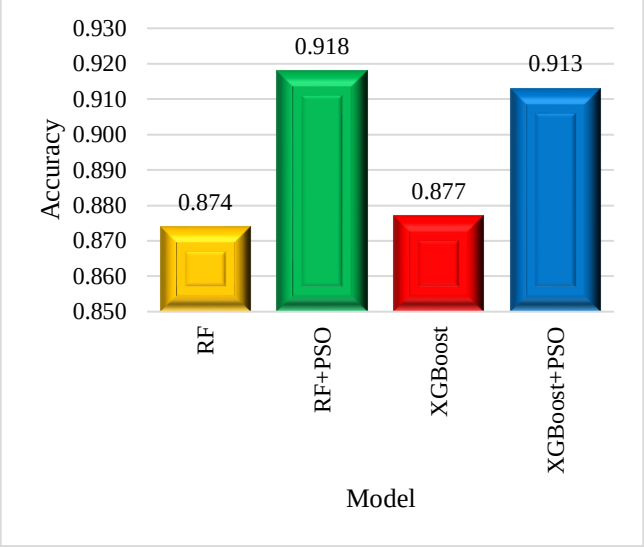


FIGURE 11. Comparison of AUC Results in Heart Disease Prediction

According to the study's findings, PSO feature selection can increase the classification accuracy of survival rates for heart disease patients when Random Forest and XGBoost are applied. This implies that the combined approach may be able to significantly enhance clinical systems' prediction performance, enabling doctors to treat patients with heart disease with more accuracy. The enhanced accuracy of the predictive model derived from the medical dataset suggests that these findings may significantly progress the application of artificial intelligence in the healthcare sector to enhance the caliber of medical services.

V. CONCLUSION

On the basis of this study, the Random Forest and XGBoost classification algorithms along with the PSO feature selection method were applied to the heart disease dataset. The optimal weight determined by PSO is intended to enhance the classification of heart disease. the conducted tests on the heart disease dataset utilizing the XGBoost (Extreme Gradient Boosting) and Random Forest classifications without the application of PSO, obtained AUC values are sequentially 0.877 and 0.874, falling within the category of sufficiently good values. Meanwhile, when employing the XGBoost (Extreme Gradient Boosting) and Random Forest classification methods with the integration of Particle Swarm Optimization (PSO), obtained AUC values are sequentially 0.913 and 0.918, which fall into the category of favorable values. The Accuracy value for heart disease with Random Forest and XGBoost rose as a result of attribute weighting in feature selection using PSO. Classifying the heart disease dataset using Random Forest with PSO and XGBoost with PSO may produce a more accurate classification than using Random Forest and XGBoost separately. By means of this, the testing before and after the implementation of PSO has the capability to enhance the performance of Extreme Gradient Boosting (XGBoost) by 0.36 and Random Forest by 0.44. PSO can increase the accuracy of the Random Forest and XGBoost approach on the heart disease dataset. It can be concluded that PSO can enhance the predictive value for the classification of expected numbers concerning heart disease patient cases.

Future studies also need to assess how well the suggested algorithm and methodology work. To guarantee that the gain in classification accuracy is the result of using the right approach, a variety of thorough evaluation measures should be used. As a result of the aforementioned advancements, future research is expected to produce more thorough and reliable findings when classifying heart disease data using a combination of Random Forest and XGBoost with PSO.

REFERENCES

[1] D. Shah, S. Patel, and S. Kumar, "Heart Disease Prediction using Machine Learning Techniques," *SN Comput. Sci.*, 2020, doi: 10.1007/s42979-020-00365-y.

[2] 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.

[3] R. Perumal and K. AC, "Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 06, pp. 4225–4234, 2020, [Online]. Available: http://sersc.org/journals/index.php/IJAST/article/view/16428

[4] M. Belgiu and L. Dra, "ISPRS Journal of Photogrammetry and Remote Sensing Random forest in remote sensing : A review of applications and future directions ˇ gut," vol. 114, pp. 24–31, 2016, doi: 10.1016/j.isprsjprs.2016.01.011.

[5] A. S. More and D. P. Rana, "Review of Random Forest Classification Techniques to Resolve Data Imbalance," pp. 72–78, 2017.

[6] R. Pavan, M. Nara, S. Gopinath, and N. Patil, "Bayesian optimization and gradient boosting to detect phishing websites," *2021 55th Annu. Conf. Inf. Sci. Syst. CISS 2021*, pp. 2–6, 2021, doi: 10.1109/CISS50987.2021.9400317.

[7] C. Chen & Guestrin, "XGBoost: A Scalable Tree Boosting System," *J. Assoc. Physicians India*, 2016, [Online]. Available: ISBN 978-1-4503-%0A4232-2/16/08

[8] O. Almomani, "SS symmetry Detection System Based on PSO , GWO , FFA and," 2020.

[9] M. Batool, A. Jalal, and K. Kim, "Sensors Technologies for Human Activity Analysis Based on SVM Optimized by PSO Algorithm," *2019 Int. Conf. Appl. Eng. Math.*, pp. 145–150, 2019.

[10] E. Prasetyo and B. Prasetiyo, "Increased Classification Accuracy C4.5 Algorithm Using Bagging Techniques in Diagnosing Heart Disease," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 5, pp. 1035–1040, 2020, doi: 10.25126/jtiik.202072379.

[11] M. Abualkibash, "U SING D IFFERENT M ACHINE L EARNING A LGORITHMS O N K DD -99 A ND N SL -K DD D ATASETS - A R EVIEW P APER," no. June 2019, 2021, doi: 10.5121/ijcsit.2019.11306.

[12] H. Tyralis and G. Papacharalampous, "Scientists and Practitioners and Their Recent History," 2019.

[13] A. E. Maxwell *et al.*, "Implementation of machine-learning classification in remote sensing : an applied review sensing : an applied review," *Int. J. Remote Sens.*, vol. 39, no. 9, pp. 2784–2817, 2018, doi: 10.1080/01431161.2018.1433343.

[14] H. B. Kibria, "The Severity Prediction of The Binary And Multi-Class Cardiovascular Disease - M ULTI -C LASS C ARDIOVASCULAR D ISEASE - A M ACHINE L EARNING -B ASED F USION A PPROACH," no. March, 2022, doi: 10.48550/arXiv.2203.04921.

[15] L. Breiman, "Random Forests. Machine Learning," *Netherlands Kluwer Acad. Publ.*, 2001.

[16] G. Biau, "Analysis of a Random Forests Model," *J. Mach. Learn. Res.*, p. Vol.49, No.5, pp. 373–381, 2012.
[17] Z. Bingzhen, "A Random Forest Classification Model for Transmission Line Image Processing," no. Iccse, pp. 613–617, 2020.
[18] B. A. Tama, L. Nkenyereye, S. M. R. Islam, and K. S. Kwak, "An enhanced anomaly detection in web traffic using a stack of classifier ensemble," *IEEE Access*, vol. 8, pp. 24120–24134, 2020, doi: 10.1109/ACCESS.2020.2969428.
[19] B. Engineering and N. Firdous, "pulmonary embolism-a non-cardiac cause of cardiac arrest Handling of derived imbalanced dataset using XGBoost for identification of pulmonary embolism — a non - cardiac cause of cardiac arrest," no. December, 2021, doi: 10.1007/s11517-021-02455-2.
[20] Y. Li, D. Yao, J. Yao, and W. Chen, "A particle swarm optimization algorithm for beam angle selection in intensity-modulated radiotherapy," vol. 3491, doi: 10.1088/0031-9155/50/15/002.
[21] M. Moodi, M. Ghazvini, and H. Moodi, "Knowledge-Based Systems A hybrid intelligent approach to detect Android Botnet using Smart," *Knowledge-Based Syst.*, vol. 222, p. 106988, 2021, doi: 10.1016/j.knosys.2021.106988.
[22] A. R. Syulistyo, D. M. J. Purnomo, M. F. Rachmadi, and A. Wibowo, "Convolutions Subsampling Convolutions Gaussian connection Full connection Full connection Subsampling," *JIKI (Jurnal Ilmu Komput. dan Informasi) UI*, vol. 9, no. 1, pp. 52–58, 2016.
[23] F. Gorunescu, *Data Mining: Concepts, models and techniques*. 2011.
[24] D. R. Chandranegara, S. Arifianto, and H. Wibowo, "Aircraft Data Analysis Using Data Void Elimination and Data Smoothing Methods," *J. POROS Tek.*, vol. 12, no. 1, pp. 1–7, 2020.
[25] I. Behravan, "An optimal SVM with feature selection using multi-objective PSO," pp. 76–81, 2016.
[26] J. Han, M. Kamber, and J. Pei, "Third Edition : Data Mining Concepts and Techniques," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2012, [Online]. Available: http://library.books24x7.com/toc.aspx?bkid=44712
[27] K. Sumathi, S. Kannan, and K. Nagarajan, "Data Mining: Analysis of student database using Classification Techniques," *Int. J. Comput. Appl.*, vol. 141, no. 8, pp. 22–27, 2016, doi: 10.5120/ijca2016909703.
[28] H. Wei, C. Hu, S. Chen, Y. Xue, and Q. Zhang, "Establishing a software defect prediction model via effective dimension reduction," *Inf. Sci. (Ny).*, vol. 477, pp. 399–409, 2019, doi: 10.1016/j.ins.2018.10.056.
[29] D. Rodriguez, I. Herraiz, R. Harrison, J. Dolado, and J. C. Riquelme, "Preliminary Comparison of Techniques for Dealing with Imbalance in Software Defect Prediction Categories and Subject Descriptors," *Proc. 18th Int. Conf. Eval. Assess. Softw. Eng. - EASE '14*, 2014.
[30] M. Ajdani and H. Ghaffary, "Introduced a new method for enhancement of intrusion detection with random forest and PSO algorithm," no. November 2020, pp. 1–10, 2021, doi: 10.1002/spy2.147.
[31] H. Jiang, Z. He, G. Ye, and H. Zhang, "Network Intrusion Detection Based on PSO-Xgboost Model," *IEEE Access*, vol. 8, pp. 58392–58401, 2020, doi: 10.1109/ACCESS.2020.2982418.

AUTHORS BIOGRAPHY

Muhammad Ridho Ansyari originated in Banjarmasin, South Kalimantan. Since 2018, he has pursued his academic endeavors as a student Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Lambung Mangkurat. His current area of research lies within the realm of Data Science. Additionally, his final project entailed conducting research that centered around the classification of Heart Disease.

Muhammad Itqan Mazdadi is a lecturer in the Department of Computer Science, Lambung Mangkurat University. His research interest is centered on Data Science and Computer Networking. Before becoming a lecturer, he completed his undergraduate program in the Computer Science Department at Lambung Mangkurat University In 2013. He then completed his master's degree from Department of Informatics at Islamic Indonesia University, Yogyakarta. Currently, he serves as the Secretary of the Computer Science Department at Lambung Mangkurat University.

Fatma Indriani is a lecturer in the Department of Computer Science, Lambung Mangkurat University. Her research interest is focused on Data Science. Before becoming a lecturer, she completed her undergraduate program in the Informatics Department at Bandung Institute of Technology. In 2008, she started working as a lecturer in the Computer Science department at Lambung Mangkurat University. She then completed her master's degree at Monash University, Australia in 2012. And her latest education is a doctorate degree in Bioinformatics at Kanazawa University, Japan, which was completed in 2022. The research fields she focuses on are Data Science and Bioinformatics.

Dwi Kartini received her bachelor's and master's degrees in computer science from the Faculty of Computer Science, Putra Indonesia University "YPTK" Padang, Indonesia. Her research interests include the applications of Artificial Intelligence and Data Mining. She is an assistant professor in the Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University in Banjarbaru, Indonesia.

Triando Hamonangan Saragih is a lecturer in Department of Computer Science, Lambung Mangkurat University. His research interest is focused on Data Science. He completed his bachelor's degree in Informatics at Brawijaya University, Malang in 2016. After that, he pursued a master's degree in Computer Science Brawijaya University, Malang in 2018. The research field he is involved in is Data Science.