International Journal of Electrical and Computer Engineering (IJECE)
Vol. 14, No. 4, August 2024, pp. 4563~4576
ISSN: 2088-8708, DOI: 10.11591/ijece.v14i4.pp4563-4576

The use of genetic algorithm and particle swarm optimization on tiered feature selection method in machine learning-based coronary heart disease diagnosis system

Wiharto[1], Yasmin Mufidah[1], Umi Salamah[1], Esti Suryani[2], Sigit Setyawan[3]
[1]Department of Informatics, Faculty of Information Technology and Data Science, Universitas Sebelas Maret, Surakarta, Indonesia
[2]Department of Data Science, Faculty of Information Technology and Data Science, Universitas Sebelas Maret, Surakarta, Indonesia
[3]Department of Medicine, Faculty of Medicine, Universitas Sebelas Maret, Surakarta, Indonesia

Article Info

Article history:

Received Oct 26, 2023
Revised Mar 7, 2024
Accepted Mar 16, 2024

Keywords:

CatBoost algorithm
Coronary heart disease
Feature selection
Genetic algorithm
Particle swarm optimization

ABSTRACT

Coronary heart disease (CHD) is a leading global cause of death. Early detection is the right step to reduce mortality rates and treatment costs. Early detection can be developed using machine learning by utilizing patient medical record datasets. Unfortunately, this dataset has excessive features which can reduce machine learning performance. For this reason, it is necessary to reduce the number of redundant features and irrelevant data to improve machine learning performance. Therefore, this research proposes a tiered of feature selection model with genetic algorithm (GA) and particle swarm optimization (PSO) to improve the performance of the diagnosis model. The feature selection model is evaluated using parameters derived from the confusion matrix and using the CatBoost machine learning algorithm. Model testing uses z-Alizadeh Sani, Cleveland, Statlog, and Hungarian datasets. The best results for this model were obtained on the z-Alizadeh Sani dataset with 6 selected features from 54 features and the resulting performance for accuracy parameters was 99.32%, specificity 98.57%, sensitivity 100.00%, area under the curve (AUC) 99.28%, and F1-Score 99.37%. The proposed feature selection model is able to provide machine learning performance in the very good category. The diagnostic model proposed is of excellent standard.

Corresponding Author:

Wiharto
Department of Informatics, Faculty of Information Technology and Data Science, Universitas Sebelas Maret
Jl. Ir Sutami No. 36A, Kentingan, Jebres, Surakarta, Indonesia
Email: wiharto@staff.uns.ac.id

1. INTRODUCTION

Coronary heart disease (CHD) arises from impaired function of the heart and blood vessels. It is a primary cause of mortality worldwide [1]. The World Health Organization (WHO) reported that CHD caused the deaths of up to 17.9 million individuals in 2019. Research by Alizadehsani *et al.* [2] indicates that 25% of individuals with CHD die unexpectedly and without any preceding symptoms. Early detection of coronary heart disease (CHD) is essential to decrease mortality rates associated with the disease [3]. Currently, CHD diagnosis is developed using machine learning models. However, the process necessitates extensive medical record data. The abundance of medical record data often results in numerous features that do not directly facilitate diagnosis [4] and can ultimately negatively influence machine learning performance. To address these issues, data mining techniques may be employed, specifically through the feature selection method. Feature selection can identify features that enhance the effectiveness of machine learning-based diagnostic models [5].

Numerous studies have been conducted on developing machine learning-based models for diagnosing coronary heart disease that uses feature selection. Kolukisa and Bakir [5] compared various feature selection models, including chi square, information gain, ReliefF, and support vector machine (SVM). The best performance was produced when using the z-Alizadeh Sani dataset with SVM feature selection producing 25 features, with an accuracy of 91.78%. Shahid and Singh [6] compared various feature selection techniques using the z-Alizadeh Sani dataset. The top-performing models were a feature selection model utilizing SVM weight and a classification model combining particle swarm optimization (PSO) with the emotional neural network (EmNN) algorithm. The resulting model achieved an accuracy of 88.34%, precision of 92.37%, sensitivity of 91.85%, specificity of 78.98%, and F1-Score of 92.12%. The research conducted by Kanagarathinam *et al.* [7] utilized Pearson's correlation feature selection technique on a dataset composed of Hungarian, Swiss, Cleveland, and Long Beach datasets. The test yielded an accuracy performance parameter of 94.34% with 10 features chosen out of a total of 13.

Several other studies have utilized computational intelligence algorithms like genetic algorithms (GA) and PSO for feature selection. Wiharto *et al.* [8] conducted research on feature selection using a combination of genetic algorithms with SVM and fast correlation based filter (FCBF) for a classification model with random forest. They evaluated their approach on z-Alizadeh Sani, Cleveland, and Statlog datasets. The study's findings revealed an accuracy rate of 94.6% and an area under the curve (AUC) rate of 97.5% based on 8 out of 54 selected features using the z-Alizadeh Sani dataset. Similarly, the Cleveland dataset showcased an 83% value after choosing 6 out of 13 featured elements. Notably, El-Shafiey *et al.* [4] examined further by employing a genetic algorithm, wherein modifications were made to the selection operator. Individuals who fail to make the selection process via the genetic algorithm will undergo processing by PSO. Individuals selected during the PSO stage will constitute the new population in the genetic algorithm. Both the genetic algorithm and PSO processes aim to optimize accuracy values via the use of random forest (RF). The proposed method was applied to the Cleveland and Statlog datasets, resulting in the best performance being achieved in the Cleveland dataset. Specifically, utilizing a subset of 7 out of 13 features, an accuracy score of 95.6% and an AUC value of 94% were obtained. This study demonstrates that employing genetic algorithms and PSO methods leads to improved performance compared to previous research that only employs a single computational intelligence algorithm.

In the CHD diagnostic system, selecting the appropriate classification model is a vital aspect in developing an effective and efficient system, in addition to the feature selection method. An accurate and reliable classification model significantly enhances the accuracy of CHD diagnosis. In the study conducted by Kanagarathinam [7], the performance of various algorithms, including naïve Bayes, XGBoost, k-nearest neighbors (kNN), SVM, multi-layer perceptron (MLP), and CatBoost, was evaluated to predict CHD on a combination of datasets from Hungarian, Switzerland, Cleveland, and Long Beach. The results showed that the CatBoost algorithm achieved the highest accuracy rate of 94.34%.

Medical data often suffers from imbalanced data, resulting in biased classification models that favour the majority class and produce false classification results. Additionally, medical data frequently has a large number of features, which must be carefully considered. To overcome these challenges, sampling methods such as oversampling and undersampling can be used, and there are several available techniques for performing sampling. Majhi and Kashyap's study [9] employed a variety of data sampling techniques, including borderline synthetic minority over-sampling technique (BD-SMOTE), support vector machine-SMOTE, adaptive synthetic sampling (ADASYN), edited nearest neighbour-SMOTE (ENN-SMOTE), and Tomek-link SMOTE, in the analysis of hospital ICU patient data. The efficacy of these techniques was tested across 12 classification models, revealing that the ENN-SMOTE sampling method in combination with the CatBoost classification algorithm produced the most favourable results in one of the datasets used.

Referring to previous research, the use of feature selection in the CHD diagnosis system, when tested using the z-Alizadeh feature dataset, resulted in 8 features, with an accuracy of 94.5%. These features are produced from the tier of methods, namely the genetic algorithm and FCBF. Unfortunately, this method when using the Cleveland dataset was only able to produce an accuracy of 83%, but the number of features produced was relatively small, namely 6 features. The feature selection model using the hybrid GA and PSO method is able to provide accuracy reaching 95.6% but requires as many as 7 features. Referring to this, the tiered approach used in hybrid GA and FCBF is very effective in reducing the number of features, compared to hybrid GA and PSO. This is because of the hybrid GA and PSO method, the feature population that is not selected in the GA selection process will be processed by PSO, while in the hybrid GA and FCBF, the features resulting from GA will be selected again by FCBF. This approach produces an optimal number of features, so this research proposes the use of GA and PSO on Tiered of feature selection (GAPSO-TFS) model in a machine learning-based CHD diagnosis system. The diagnostic system also employs the CatBoost classification algorithm and implements the ENN-SMOTE sampling method to balance the data. To evaluate the system, k-fold cross validation is conducted, and performance is measured using a confusion matrix. The

study measured Accuracy, Sensitivity, Specificity, AUC, and F1-Score performance parameters. The testing datasets selected were from the University of California (UCI), Irvine, machine learning repository [10]–[13], specifically z-Alizadeh Sani, Statlog, Cleveland, and Hungarian datasets.

2. METHOD

This study presents a machine learning-based model for diagnosing coronary heart disease. The model incorporates tiered feature selection through GA and PSO methods, known as GAPSO-TFS. The research method comprises six stages: data collection, data preprocessing, feature selection, data sampling, classification, and result evaluation. A detailed representation of the entire method is included in Figure 1.

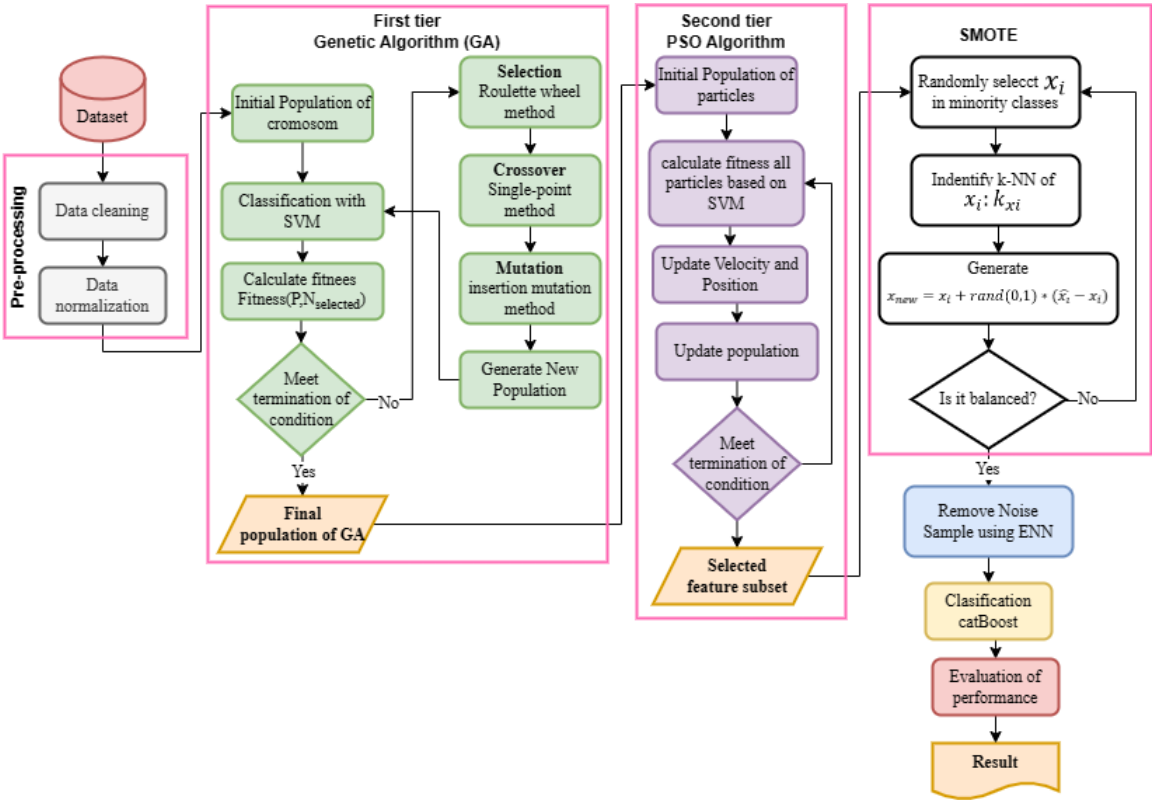


Figure 1. Proposed model

2.1. Material

The study will utilize the z-Alizadeh Sani, Statlog, Cleveland, and Hungarian datasets obtained from UCI machine learning [10]–[13]. The z-Alizadeh Sani dataset provides extensive features, containing heart disease data from 303 patients with 54 attributes. The Statlog dataset includes heart disease data from 261 patients with 13 attributes. The Cleveland and Hungarian datasets belong to a group of heart disease datasets. This dataset comprises of 76 attributes. However, usually only 13 attributes with complete attribute values are utilized in research. The Cleveland dataset encompasses heart disease data from 297 patients, while the Hungarian dataset encompasses heart disease data from 294 patients. Table 1 depicts a comparison of the amount of data for positive and negative CHD bees.

Table 1. Heart disease dataset

Dataset	Number of Attributes	Number of data	Positive and negative ratio
z-Alizadeh Sani	54	303	1: 2.5
Statlog	13	261	1: 0.78
Cleveland	13	303	1: 0.85
Hungarian	13	294	1: 0.56

2.2. Pre-processing

The preprocessing stage cleans the data for feature selection. It involves checking for empty attribute values, encoding category data, and performing data normalization. For empty-valued attributes, if the attribute data is continuous, it is replaced with the average value, but if it is discrete, it is replaced with the mode value. Data normalization is necessary because the dataset used comprises various and extensive value ranges. The data variations are converted to a value range of 0-1 to enhance performance and speed in the classification process. The Min Max normalization method is employed for data normalization in this study.

2.3. Feature selection

The feature selection stage is a natural extension of the pre-processing stage in machine learning. Its purpose is to select the optimal features to achieve optimal classification performance. In this research, we employed the wrapper method of feature selection. This method employs genetic algorithms and PSO to identify the optimal subset of features. The selection process is done in stages to ensure a balanced selection of features that adequately represent the data. First, feature selection is conducted using a genetic algorithm. The outcomes of the genetic algorithm feature selection establish the foundational population for the initial population of the feature selection process utilizing the PSO algorithm. The GAPSO-TFS method for feature selection can be visualized in Figure 1.

2.3.1. The first tier: genetic algorithm

The genetic algorithm is a search algorithm that adapts the natural selection process of living things and genetics [14]. Its objective is to find the global optimum value of a problem while maintaining the best solution in each generation. As a result, each generation gradually moves towards a better solution [15]. In genetic algorithm problem-solving, a chromosome represents each solution. Each chromosome contains genes that represent selected features as either 1 or 0, depicted in Figure 2. The chromosomes undergo evaluation utilizing the SVM classification algorithm, and the accuracy outcomes calculate the fitness function. Equation (1) illustrates the minimum expected fitness value. The genetic algorithm's feature selection process is marked in Figure 1, in the first tier.

$$Fitness\ value(P, N_{selected}) = \alpha \times (1 - P) + (1 - \alpha) \times \frac{N_{selected}}{N_{features}} \quad (1)$$

The parameter α, accuracy value of the classification model (P), number of selected features ($N_{selected}$), and total number of features from the dataset ($N_{features}$) will be taken into account. The value of α utilized for this research is 0.99. The selection process in this research will utilize the roulette wheel method to choose the best chromosome for producing a new chromosome. Next, the crossover stage combines two selected chromosomes to produce offspring. For this study, a single point technique was used for the crossover stage. The final stage involves mutation, which randomly changes gene values to increase individual diversity [14]. This process utilizes the insertion mutation technique. The complete feature selection process with GA is shown in Figure 1.

2.3.2. The second tier: particle swarm optimization

Particle swarm optimization (PSO) represents a mode of swarm intelligence (SI) optimization algorithm inspired by the social behavior of animals, like birds, fish, ants, bees, and termites [16]. According to PSO, each solution is depicted as a particle in a swarm constituting a potential solution to the D-dimensional space problem. Referring to this concept, in this research, a particle is a collection of features whose value is 0 or 1, if it is 1, it means that the feature is selected, and if 0, the feature is not selected. The particle modeling is shown in Figure 3, while the feature selection process is shown in Figure 1 in the second tier. This process begins by initializing the particle with a randomly initialized velocity and initial position. The position in the search space is represented in a vector [17], as shown in (2).

Feature dataset	Feature-1	Feature-2	Feature-3	.....	Feature-n
Chromosom	1	0	1	.....	0

Figure 2. Feature-chromosome model

Feature dataset	Feature-1	Feature-2	Feature-3	.....	Feature-n
Particle	1	0	1	.....	0

Figure 3. Feature-particle model

$$X_i = (x_{i1}, x_{i2}, \dots\dots\dots\dots, x_{iD}) \quad (2)$$

Each particle also retains the previous best position represented in the vector, as illustrated in (3).

$$P_i = (p_{i1}, p_{i2}, \dots\dots\dots\dots, p_{iD}) \quad (3)$$

The particles within the swarm move throughout the search space to locate the optimal solution. Thus, each particle possesses a velocity portrayed by a vector, illustrated in (4).

$$V_i = (v_{i1}, v_{i2}, \dots\dots\dots\dots, v_{iD}) \quad (4)$$

To choose individual particles, the classification model uses SVM, and the fitness function reduces the (1). PSO searches for the best solution by updating each particle's velocity and position, based on (2) and (3), which are written as (5) and (6) [17].

$$v_{id} = w * v_{id} + c_1 * r_1 * (P_{id} - x_{id}) + c_2 * r_2 * \left(P_{gd} - x_{id}\right) \quad (5)$$

$$x_{id} = x_{id} + v_{id} \quad (6)$$

where v_{id} represents particle velocity, x_{id} represents the current particle position, w represents moment of inertia, $c_1$represents cognitive coefficient that accounts for individual behavior, while c_2 represents social coefficient that accounts for group behavior. The value for both parameters c_1 and c_2 is 2. Parameter P_{id} represents personal best (pbest), P_{gd} represents global best (gbest), while $r_1$and r_2 represent random numbers ranging from 0 to 1.

2.4. Data sampling

Data sampling is utilized to rectify data imbalance by increasing or decreasing the amount of data. The CHD dataset exhibits a notable difference between patients with normal labels and patients with heart disease labels. Thus, this study will implement data sampling techniques where the minority class is oversampled and the majority class is undersampled using the SMOTE method, to overcome this issue.

The SMOTE technique is a conventional approach to oversampling in the minority class, leveraging the k-nearest neighbor (K-NN) algorithm [18]. By using linear interpolation to balance the dataset, the algorithm increases the number of samples in the minority class. The method involves picking $S_{j\ min}$ samples from the k nearest neighbors for every minority class $S_{i\ min}$, resulting in synthesizing a new minority class sample using (7).

$$S_{new} = S_{i\ min} + rand(0,1)(S_{j\ min} + S_{\mathrm{i\ min}}), i = 1,2,\dots,n,\ j = 1,2,\dots,k \quad (7)$$

The SMOTE algorithm presents limitations in addressing problems of sample overlap and data noise [19]. To mitigate these limitations, Batista *et al.* [20] introduced the SMOTE-ENN algorithm, a combination of SMOTE and the edited nearest neighbors (ENN) algorithm. ENN is employed as an undersampling method to enhance the classification performance of minority samples by decreasing the majority samples. This method is modeled on the SMOTE algorithm, which is susceptible to noisy data [18]. To enhance the quality of the resulting data, data from the majority class, which is deemed noisy, is reduced. The algorithm examines each majority class datum in relation to its nearest majority and minority class. If the majority class is closer to its minority class neighbor, the data is flagged as noise and eliminated from the dataset. This ENN algorithm combination aims to remove noise data generated by the SMOTE algorithm to produce higher quality data [19].

2.5. Classification model

The feature selection process results will be entered into the classification process to generate a positive or negative decision regarding CHD. The k-fold cross-validation technique will be employed at this stage, dividing the data into almost equal parts through random division. The data division results will be used as training and testing data to build the diagnosis system model, with the data classified into two labels: normal and heart disease patients. In this study, we will use the CatBoost method as our classification approach. CatBoost is one of the algorithms under the gradient boosting decision tree (GBDT) [21]. This technique introduces two innovations related to ordering, namely ordered target statistics and ordered boosting.

One issue with utilizing the gradient boosting algorithm is that the learned model's distribution can shift due to the repeated boosting process. To avoid target leakage, this ordered boosting technique can facilitate the learning process. Additionally, a common challenge is managing attributes that have categorical data types. This algorithm enables conversion of categories to numerical values without the initial data processing stage, using the ordered target statistics method [22]. The decision tree construction stage is critical to CatBoost. The algorithm utilizes a two-step decision tree building process involving selection of a tree structure based on the GBDT algorithm and assignment of leaf values. At each iteration, the construction of every tree is achieved by evaluating the loss reduction of the prior tree. The fundamental predictor employed in the CatBoost algorithm is a balanced oblivious decision tree (ODT) or symmetry tree, which is immune to overfitting and capable of accelerating the testing process [22].

2.5. Result evaluation

The CatBoost method is used to evaluate the performance of classification results in the result evaluation stage. The evaluation process involves calculating the confusion matrix value utilizing the confusion matrix table. The confusion matrix provides four values that indicate the model's classification results. True positive (TP) represents the number of positive data points detected accurately by the model. True negative (TN) refers to the number of accurately identified negative data by the model. False negative (FN) is the number of positive data incorrectly identified as negative. False positive (FP) is the number of negative data incorrectly identified as positive. A comprehensive description is provided in Figure 4 [23]. Based on the confusion matrix, accuracy, specificity, sensitivity, and AUC performance parameters can be calculated using (8)-(12).

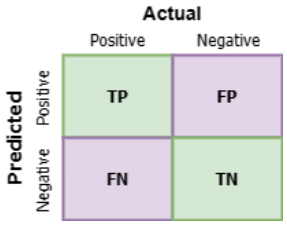


Figure 4. Confusion matrix

$$\text{Accuracy} = \text{ACC} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})} \times 100\% \quad (8)$$

$$\text{Specificity} = \text{SEN} = \frac{\text{TN}}{(\text{TN}+\text{FP})} \times 100\% \quad (9)$$

$$\text{Sensitivity} = \text{SPE} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \times 100\% \quad (10)$$

$$\text{AUC} = \frac{1}{2}\left(\frac{\text{TP}}{(\text{TP}+\text{FN})} + \frac{\text{TN}}{(\text{TN}+\text{FP})}\right) \times 100\% \quad (11)$$

$$\text{F1} - \text{Score} = \frac{2 \times \text{TP}}{(2 \times \text{TP}+\text{FP}+\text{FN})} \times 100\% \quad (12)$$

3. RESULTS AND DISCUSSION

3.1. Data pre-processing results

The pre-processing stages consisted of encoding category data, evaluating empty/null values, and detecting data anomalies [24]. Subsequently, empty/null values were examined, and it was determined that the z-Alizadeh Sani and Statlog datasets contained none. The Cleveland dataset, however, had six empty values, which were removed from the dataset. For the Hungarian dataset, some variables have missing values. In order to deal with this issue, the missing values are replaced with the average value for the *trestbps*, *chol*, *thalach*, *fbs*, *exang*, and *restecg* variables because they have continuous values. In order to deal with this issue, the missing values are replaced with the average value for the *trestbps*, *chol*, *thalach*, *fbs*, *exang*, and *restecg* variables because they have continuous values. This replacement of missing values with averages is done to optimize the modeling process. The *fbs*, *exang*, and *restecg* variables also have missing

values. However, since the data for the attribute is discrete, any missing values are replaced with the mode value for each attribute. Additionally, the attributes for *slope, ca,* and *thal* have over 50% of their data with missing values, resulting in a significant amount of inaccurate data for these three attributes. Therefore, we removed the aforementioned three attributes from the Hungarian dataset. Our subsequent procedure encompasses normalizing all datasets using the min-max normalization method. Its outcome converts the data into decimal numbers within the 0-1 range.

3.2. Feature selection results

The pre-processed data undergoes feature selection using three scenarios. The first scenario entails the use of a GA. The second scenario involves feature selection using the PSO algorithm. The third scenario was conducted with the GA and PSO tiered feature selection (GAPSO-TFS) method. The GA parameters for this test utilized a crossover probability (Pc) of 0.8 and a mutation probability (Pm) of 0.01. For every tested technique, the process utilized 100 particles. The GA method used 100 generations, and the PSO method utilized 100 iterations. The GA cascading method employed 20 generations, whereas the PSO used 20 iterations.

Table 2 displays the findings of feature selection testing for the four datasets utilizing the specified parameters. Through analysis of the obtained test results, we were able to reduce the number of features significantly during feature selection. Among the three scenarios, the GAPSO-TFS tiered feature selection method achieved the most feature reduction. Furthermore, the GAPSO-TFS method required fewer generations than the GA and PSO methods. This demonstrates that the GAPSO-TFS method's feature selection technique is more efficient than other techniques, as it requires fewer generations and results in a smaller number of selected features.

Table 2. Feature selection testing results

Dataset	Method	#Feature	Name of selected features
z-Alizadeh Sani	GA	14	'HTN', 'FH', 'CRF', 'DLP', 'PR', 'Diastolic Murmur', 'Typical Chest Pain', 'Q Wave', 'CR', 'BUN', 'HB', 'Lymph', 'EF-TTE', 'Region RWMA'
	PSO	18	'DM', 'Current Smoker', 'CRF', 'Edema', 'Weak Peripheral Pulse', 'Systolic Murmur', 'Diastolic Murmur', 'Typical Chest Pain', 'LVH', 'FBS', 'CR', 'BUN', 'HB', 'K', 'Na', 'Neut', 'EF-TTE'
	GAPSO-TFS	6	'BMI', 'Current Smoker', 'FH', 'Obesity', 'CVA', 'PR'
Cleveland	GA	7	'age', 'cp', 'chol', 'thalach', 'exang', 'oldpeak', 'slope'
	PSO	7	'sex', 'fbs', 'restecg', 'thalach', 'exang', 'slope', 'ca'
	GAPSO-TFS	4	'age', 'sex', 'cp', 'trestbps'
Statlog	GA	5	'age', 'sex', 'restecg', 'thalach', 'slope'
	PSO	5	'sex', 'cp', 'fbs', 'slope', 'ca'
	GAPSO-TFS	4	'age', 'trestbps', 'chol', 'fbs'
Hungarian	GA	5	'age', 'sex', 'chol', 'restecg', 'exang'
	PSO	5	'age', 'sex', 'fbs', 'restecg', 'exang'
	GAPSO-TFS	3	'age', 'cp', 'trestbps'

In the z-Alizadeh Sani dataset, the GAPSO-TFS method identifies the 6 most significant features, namely body mass index (BMI), current smoker (indicating the smoking status of the patient), Family History (FH) indicating whether the family has a history of heart disease, Obesity (indicating whether the patient suffers from obesity), cerebrovascular accident (CVA) indicating functional brain disorders, and pulse rate (PR) indicating the heart rate of the patient). The use of these features can aid in predicting heart disease risk. In the Cleveland database, four attributes have been chosen: Age for age, sex for gender, chest pain type (Cp) for the type of chest pain felt by the patient, and resting blood pressure (Trestbps) for blood pressure after resting. In the Statlog dataset, four attributes were selected, namely age denoting age, Trestbps denoting blood pressure after rest, Chol denoting the quantity of cholesterol in the blood, and fasting blood sugar (Fbs) denoting levels of blood sugar after fasting. Concerning the Hungarian dataset, three characteristics were chosen: age represents age, Cp denotes the kind of chest pain the patient experienced, and Trestbps indicates the blood pressure after resting.

3.3. Classification results

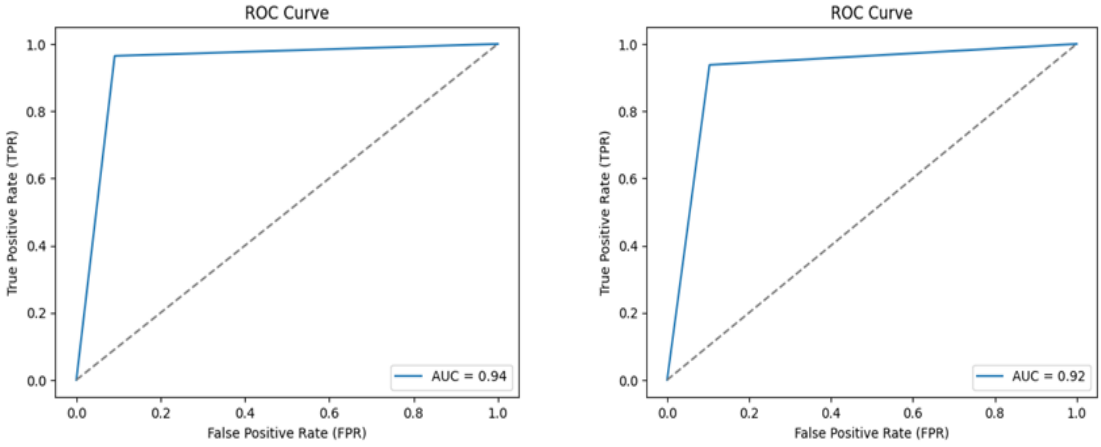
The final stage of testing assesses the classification outcomes post-feature selection. To balance the data for classification, we leveraged the SMOTE-ENN technique. We employed the CatBoost algorithm for classification. The accuracy, specificity, sensitivity, AUC, and F1-Score represent the parameters used to assess the classification performance, calculated using (8)-(12). During the division of training data and test data, we utilize the k-fold cross-validation method with k=10. The dataset is split into 10 sections, with each section taking a turn as the test data while the remaining nine sections serve as the training data. This process is then alternated to ensure that every data point is used for both test and training purposes.

3.3.1. Evaluation of z-Alizadeh Sani dataset results

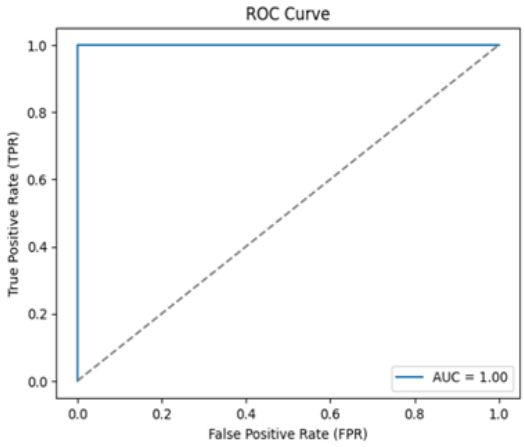
The classification model underwent testing on the z-Alizadeh Sani dataset using features selected from each feature selection method. Table 3 presents the test results, calculated by averaging the 10-fold cross validation method. Based on these outcomes, it is demonstrated that the GAPSO-TFS method optimally reduces 54 features to 6 selected features, performing better than either GA or PSO feature selection. The GAPSO-TFS model reduces features by 88.89%, which improves performance. The resulting AUC value falls into the excellent category, exceeding 90%. Figure 5 displays the receiver operating characteristic (ROC) curve of the AUC value. The three graphs illustrate the ROC curve of the z-Alizadeh Sani dataset test results, using the 10-fold cross-validation approach. Figure 5(a) shows that the GA-based feature selection model produces a curve that goes to the upper left corner, which means it shows better performance. The same thing for the feature selection model using PSO shown in Figure 5(b), where if we look at the AUC value based on Table 3, there is only a difference of 0.95%. This condition is much different when compared to the GAPSO-TFS model, where the ROC curve shown in Figure 5(c), the curve is closer to the upper left corner than the curve from GA and PSO, or further away from the diagonal line, which means the performance of the model is getting better. When compared to the GA or PSO models, the GAPSO-TFS model performs much better.

Table 3. Evaluation of z-Alizadeh Sani dataset results

Method	#Feature	ACC	AUC	SPE	SEN	F1-Score
-	54	92.50	92.33	89.67	94.77	93.32
GA	14	95.09	95.02	92.67	97.37	95.28
PSO	17	96.04	95.97	93.86	98.08	96.26
GAPSO-TFS	6	99.32	99.28	98.57	100.00	99.37



(a) (b)



(c)

Figure 5. ROCcurve of z-Alizadeh Sani with (a) GA, (b) PSO, and (c) GAPSO-TFS methods

3.3.2. Evaluation of Cleveland dataset results

Testing the classification model using the Cleveland dataset utilized the features selected in each of the feature selection methods tested. Table 4 depicts the test results, derived from the average of the 10-fold cross-validation. It is demonstrated in Table 4 that GAPSO-TFS method could decrease the attribute from 13 features to 4 features, leading to enhance the performance, as the accuracy parameter can improve over 5%. Compared to the GA and PSO methods, the GAPSO-TFS has the ability to reduce the number of features by 61.54%, specifically from 7 features down to 4 features. Displayed in Figure 6, the ROC curve illustrates the AUC value, which amounts to 95.09%. The AUC value falls under the outstanding category. Figure 6(a) shows the ROC curve of the feature selection performance of GA, where the curve is closer to the upper left corner compared to Figure 6(b) which uses the PSO method. This shows that the performance of GA is better than PSO. Figure 6(c) shows the performance for the GAPSO-TFS feature selection method. The GAPSO-TFS method when compared to GA and PSO, the ROC curve is closer to the upper left corner, meaning that the performance of GAPSO-TFS is better than GA and PSO.

Table 4. Evaluation of Cleveland dataset results

Method	#Feature	ACC	AUC	SPE	SEN	F1-Score
-	13	90.00	90.01	85.00	95.18	90.85
GA	7	94.76	94.86	94.54	95.18	94.67
PSO	7	94.31	94.27	95.36	93.18	94.10
GAPSO-TFS	4	95.05	95.09	91.86	98.33	95.25

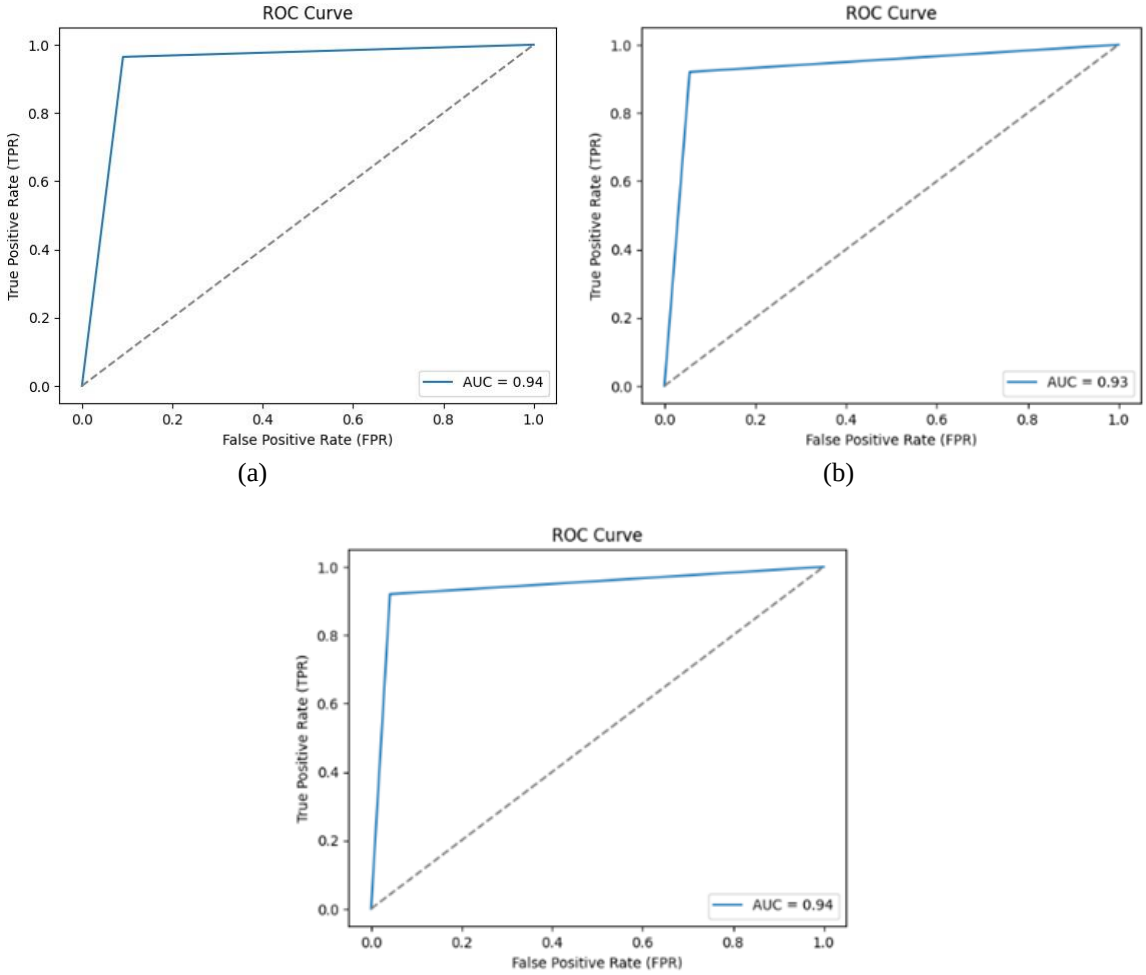


Figure 6. Cleveland's ROC curve with (a) GA, (b) PSO, and (c) GAPSO-TFS methods

3.3.3. Evaluation of Statlog dataset results

The test outcomes of the classification model on the Statlog dataset are displayed in Table 5, which is based on an average 10-fold cross-validation. According to the results, the GAPSO-TFS technique operates optimally by reducing the features from thirteen to four and exhibiting improved performance compared to GA and PSO feature selection, at 76.92% reduction rate. The ROC chart reveals an AUC value of 95.00% for the performance parameter AUC. Figure 7 demonstrates the test results for all feature selection methods on the complete ROC curve. Figure 7(a) is the ROC curve of the feature selection results with the GA method. The curve shows that the curve is further away from the diagonal line, so it is closer to the upper left corner. The same thing is also shown in Figure 7(b), which is the performance of PSO feature selection, and Figure 7(c) which is the performance of GAPSO-TFS feature selection. The advantage of the GAPSO-TFS model is the smaller number of features compared to the GA and PSO methods.

Table 5. Evaluation of Statlog dataset results

Method	#Feature	ACC	AUC	SPE	SEN	F1-Score
-	13	92.47	92.44	88.89	95.00	92.65
GA	5	93.66	93.54	93.33	93.75	93.20
PSO	5	94.45	94.44	94.89	94.00	94.36
GAPSO-TFS	4	94.95	95.00	95.00	95.00	94.90

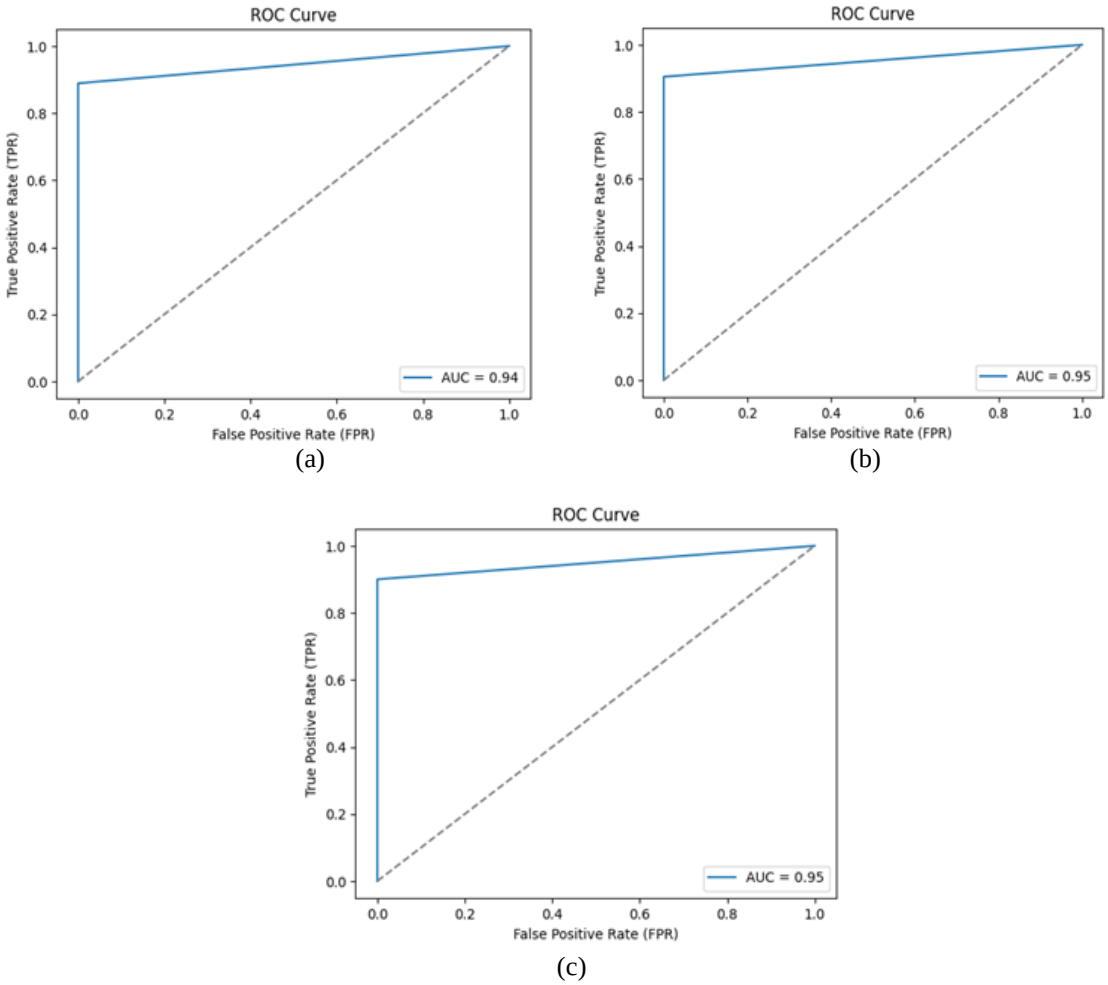


Figure 7. ROC curve of Statlog with (a) GA, (b) PSO, and (c) GAPSO-TFS methods

3.3.4. Evaluation of Hungarian dataset results

The findings of the study, which utilized GA, PSO, and GAPSO-TFS feature selection on the Hungarian dataset, have been presented in Table 6. The test outcomes were computed from an average of

10-fold cross validation. The results illustrate that the GAPSO-TFS approach performs optimally by diminishing the features from 13 to 3, a reduction of almost 75%. The results illustrate that the GAPSO-TFS approach performs optimally by diminishing the features from 13 to 3, a reduction of almost 75%. The results illustrate that the GAPSO-TFS approach performs optimally by diminishing the features from 13 to 3, a reduction of almost 75%. Despite the substantial reduction in features, the accuracy parameter improves by nearly 3%. Compared to utilizing GA and PSO feature selection methods, the GAPSO-TFS method resulted in a reduction percentage of 76.92%. Figure 8 illustrates the ROC curve that shows the AUC value. The three graphs exhibit the ROC graph of the Hungarian dataset test outcomes, excluding 10-fold cross-validation.

Table 6. Evaluation of Hungarian dataset results

Method	#Feature	ACC	AUC	SPE	SEN	F1-Score
-	13	92.69	92.77	91.52	94.01	92.12
GA	5	91.37	91.34	91.65	91.02	91.20
PSO	5	93.70	93.68	94.29	93.08	93.37
GAPSO-TFS	3	95.59	95.49	95.71	95.28	94.33

Figure 8(a) shows a curve that moves away from the top left point, namely towards the bottom of the true positive rate axis, while for Figure 8(b), the same as Figure 8(a), it is closer to the top left corner point. This shows that the performance of feature selection using PSO and GAPSO-TFS is better than GA. If you compare Figure 8(b) with Figure 8(c), then Figure 8(c) is closer to the top left corner point, which means the performance produced by GAPSO-TFS is better than PSO.

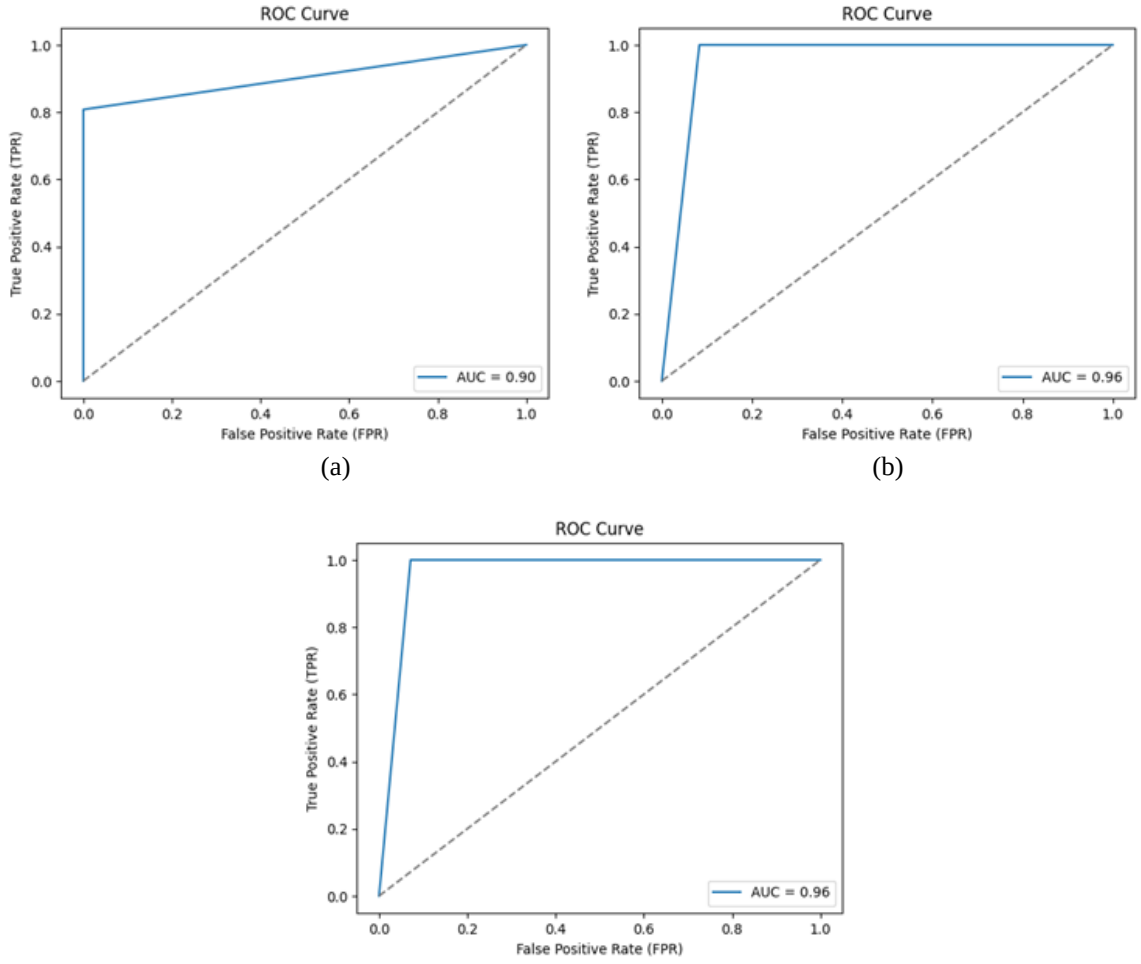


Figure 8. ROC curve of Hungarian with (a) GA, (b) PSO, dan (c) GAPSO-TFS methods

3.4. Discussion

The proposed model for selecting features has a superior ability to reduce features compared to the genetic algorithm or particle swarm optimization methods. Moreover, the high feature reduction ability results in an improved performance output. Table 7 displays the percentage of feature reduction. The average percentage of feature reduction across datasets utilizing the GAPSO-TFS method reached 76.07%. The z-Alizadeh Sani dataset demonstrated the greatest reduction at 88.89%, whereas the clevelands dataset had the lowest reduction at 61.54%.

Table 7. Percentage of feature reduction of GA-PSO model

Method	%Feature reduction				
	z-Alizadeh Sani	Clevelands	Statlog	Hungarian	Mean
GA	68.52	38.46	61.54	61.54	57.52
PSO	66.67	46.15	69.23	69.23	62.82
GAPSO-TFS	88.89	61.54	76.92	76.92	76.07

Several earlier research have proposed feature selection methods utilizing diverse classification techniques. In comparison to earlier studies, the number of features generated from previous research is evaluated, and later classification is done employing the CatBoost algorithm. Table 8 displays the performance produced by numerous previous studies. According to Table 8, the utilization of the GAPSO-TFS approach is superior in executing feature selection. The test results demonstrate that this method provides better performance. When compared to previous research [6] that utilized the z-Alizadeh Sani dataset, the proposed study displays inferior performance with a selection of only 0.08%. However, when considering the number of features required to achieve this performance level, the proposed study outperforms previous research as it only requires 6 features, while the previous study necessitated 22 features. In study 4, the use of the Statlog dataset results in improved performance, with a difference in accuracy of 0.85%. However, it requires 9 features in total, whereas the proposed model only requires 4. The system model proposed exhibits superior performance in terms of accuracy and the number of features generated across all datasets, compared to the research conducted in [8] and [25].

Table 8. Comparison of results with prior studies

Ref	Method	Dataset							
		z-Alizadeh Sani		Cleveland		Statlog		Hungarian	
		Feature	ACC	Feature	ACC	Feature	ACC	Feature	ACC
[4]	Hybrid GA PSO-RF	-	-	7	94.6	9	95.8	-	-
[6]	Weight by SVM	22	99.4	7	90.0	8	92.6	-	-
[8]	SVM-GA+FCBF	8	99.1	10	92.8	6	93.7	-	-
[25]	$\chi 2$ statistical-DNN	-	-	11	94.5	-	-	-	-

4. CONCLUSION

The heart disease diagnosis system, which utilizes the GAPSO-TFS method for feature selection and the CatBoost classification algorithm, demonstrates strong performance. The GAPSO-TFS algorithm is determined to be effective at reducing the number of features and improving evaluation performance based on testing with GA, PSO, and GAPSO-TFS. On the z-Alizadeh Sani dataset, the model achieved optimal results with only 6 features selected from the initial 54. The achieved results showed 99.32% accuracy, 98.57% specificity, 100.00% sensitivity, 99.28% AUC, and 99.37% F1-Score. The machine learning-based coronary heart disease diagnosis system model, with the GAPSO-TFS feature selection method, is categorized as excellent considering the best evaluation performance achieved on the z-Alizadeh Sani dataset. The AUC parameter provided values above 90% for all datasets.

ACKNOWLEDGEMENTS

We thank the National Research and Innovation Agency of the Republic of Indonesia, which provided research funding under the Basic Research Grant scheme under Contract No. 380.1/UN27.22/PT.01.03/2023. In addition, we would like to thank the Faculty of Information Technology and Data Science at Universitas Sebelas Maret, which has provided computer laboratory facilities, so that this research can be carried out.

REFERENCES

[1] 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.

[2] R. Alizadehsani *et al.*, "A data mining approach for diagnosis of coronary artery disease," *Computer Methods and Programs in Biomedicine*, vol. 111, no. 1, pp. 52–61, Jul. 2013, doi: 10.1016/j.cmpb.2013.03.004.

[3] C. Pan, A. Poddar, R. Mukherjee, and A. K. Ray, "Impact of categorical and numerical features in ensemble machine learning frameworks for heart disease prediction," *Biomedical Signal Processing and Control*, vol. 76, Jul. 2022, doi: 10.1016/j.bspc.2022.103666.

[4] M. G. El-Shafiey, A. Hagag, E.-S. A. El-Dahshan, and M. A. Ismail, "A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest," *Multimedia Tools and Applications*, vol. 81, no. 13, pp. 18155–18179, May 2022, doi: 10.1007/s11042-022-12425-x.

[5] B. Kolukisa and B. Bakir-Gungor, "Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis," *Computer Standards & Interfaces*, vol. 84, Mar. 2023, doi: 10.1016/j.csi.2022.103706.

[6] A. H. Shahid and M. P. Singh, "A novel approach for coronary artery disease diagnosis using hybrid particle swarm optimization based emotional neural network," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 4, pp. 1568–1585, Oct. 2020, doi: 10.1016/j.bbe.2020.09.005.

[7] K. Kanagarathinam, D. Sankaran, and R. Manikandan, "Machine learning-based risk prediction model for cardiovascular disease using a hybrid dataset," *Data & Knowledge Engineering*, vol. 140, Jul. 2022, doi: 10.1016/j.datak.2022.102042.

[8] W. Wiharto, E. Suryani, S. Setyawan, and B. P. Putra, "Hybrid feature selection method based on genetic algorithm for the diagnosis of coronary heart disease," *Journal of information and communication convergence engineering*, vol. 20, no. 1, pp. 31–40, 2022.

[9] B. Majhi and A. Kashyap, "Wavelet based ensemble models for early mortality prediction using imbalance ICU big data," *Smart Health*, vol. 28, Jun. 2023, doi: 10.1016/j.smhl.2023.100374.

[10] UCI, "UCI machine learning repository: Z-Alizadeh Sani data set," https://archive.ics.uci.edu/dataset/412/z+alizadeh+sani (accessed Nov. 24, 2022).

[11] UCI, "UCI machine learning repository: heart disease data set," https://archive.ics.uci.edu/ml/datasets/heart+disease (accessed Nov. 24, 2022).

[12] UCI, "UCI machine learning repository: Statlog (heart) data set," https://archive.ics.uci.edu/ml/datasets/statlog+(heart) (accessed Nov. 24, 2022).

[13] D. Dua and G. Casey, "UCI machine learning repository," 2017, [Online]. Available: http://archive.ics.uci.edu/ml.

[14] P. Ghamisi and J. A. Benediktsson, "Feature selection based on hybridization of genetic algorithm and particle swarm optimization," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 2, pp. 309–313, Feb. 2015, doi: 10.1109/LGRS.2014.2337320.

[15] R. R. Rajammal, S. Mirjalili, G. Ekambaram, and N. Palanisamy, "Binary grey wolf optimizer with mutation and adaptive k-nearest neighbour for feature selection in Parkinson's disease diagnosis," *Knowledge-Based Systems*, vol. 246, Jun. 2022, doi: 10.1016/j.knosys.2022.108701.

[16] F. Marini and B. Walczak, "Particle swarm optimization (PSO). a tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 149, pp. 153–165, Dec. 2015, doi: 10.1016/j.chemolab.2015.08.020.

[17] K. Premalatha and A. M. Natarajan, "Hybrid PSO and GA for global maximization," *International Journal of Open Problems in Computer Science and Mathematics*, vol. 2, no. 4, pp. 597–608, 2009.

[18] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on random forest for medical imbalanced data," *Journal of Biomedical Informatics*, vol. 107, Jul. 2020, doi: 10.1016/j.jbi.2020.103465.

[19] J. Wang, "Prediction of postoperative recovery in patients with acoustic neuroma using machine learning and SMOTE-ENN techniques," *Mathematical Biosciences and Engineering*, vol. 19, no. 10, pp. 10407–10423, 2022, doi: 10.3934/mbe.2022487.

[20] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.

[21] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *Journal of Big Data*, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-020-00369-8.

[22] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Advances in neural information processing systems*, 2018.

[23] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*. Morgan Kaufmann, 2012.

[24] K. Lohita, A. Amitha Sree, D. Poojitha, T. Renuga Devi, and A. Umamakeswari, "Performance analysis of various data mining techniques in the prediction of heart disease," *Indian Journal of Science and Technology*, vol. 8, no. 35, Dec. 2015, doi: 10.17485/ijst/2015/v8i35/87458.

[25] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, "An automated diagnostic system for heart disease prediction based on χ2 statistical model and optimally configured deep neural network," *IEEE Access*, vol. 7, pp. 34938–34945, 2019, doi: 10.1109/ACCESS.2019.2904800.

BIOGRAPHIES OF AUTHORS

Wiharto received obtained a bachelor's degree in electrical engineering (B.E.) from Universitas Telkom, Indonesia, in 1999. He obtained a master's degree in computer science from Universitas Gadjah Mada, Indonesia, in 2004 and a doctoral degree from the same university, in 2017. Currently he works as a lecturer in the Informatics Department, Faculty of Information Technology and Data Science, Universitas Sebelas Maret, Surakarta, Indonesia. His experience and areas of interest focus on artificial intelligence, clinical decision support systems, machine learning, and expert systems. He can be contacted at email: wiharto@staff.uns.ac.id.

Yasmin Mufidah received the bachelor's degree in informatics from the Faculty of Information Technology and Data Science, Universitas Sebelas Maret, Surakarta, Indonesia, in 2023. His research interests include deep learning, image processing, artificial intelligence, machine learning, and computational intelligence. She can be contacted at email: yasminmufidah@student.uns.ac.id.

Umi Salamah received her bachelor's degree from the Department of Mathematics, Universitas Sebelas Maret, Indonesia, in 1994. She received her Master's and Doctoral Degrees from the Department of Informatics Engineering, Institut Teknologi Sepuluh Nopember, Indonesia, in 2002 and 2018, respectively. Her research interests include fuzzy logic and systems, image processing, applied mathematics, and computational sciences. She can be contacted at email: umisalamah@staff.uns.ac.id.

Esti Suryani received obtained a Bachelor of Science (B.S.) from Universitas Gadjah Mada, Indonesia, 2002 and master's degree in computer science (M.Cs.) from Universitas Gadjah Mada, Indonesia, 2006. She is presently working as an assistant professor in the Department of Data Science, Faculty of Information Technology and Data Science, Universitas Sebelas Maret, Surakarta, Indonesia. His experience and areas of interest focus on image processing, fuzzy logic, data mining, and expert systems. She can be contacted at email: estisuryani@staff.uns.ac.id.

Sigit Setyawan received the medical education (M.D.) degree from the Faculty of Medicine, Universitas Sebelas Maret, Surakarta, Indonesia, in 2007, and the master's degree in tropical medicine from Universitas Gadjah Mada, Yogyakarta, Indonesia, in 2015. He is currently working as an assistant professor with the Faculty of Medicine, Universitas Sebelas Maret. His research interests include molecular biology, genomics, and health informatics. He can be contacted at email: sigitsetyawan@staff.uns.ac.id.