Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 1906466, 12 pages
https://doi.org/10.1155/2022/1906466

Research Article

An Efficient Machine Learning Model Based on Improved Features Selections for Early and Accurate Heart Disease Predication

Farhat Ullah,[1] **Xin Chen**,[1] **Khairan Rajab**,[2] **Mana Saleh Al Reshan**,[2] **Asadullah Shaikh**,[2] **Muhammad Abul Hassan**,[3] **Muhammad Rizwan**,[4] **and Monika Davidekova**[5]

[1]*School of Automation, China University of Geosciences, Wuhan 430074, China*
[2]*College of Computer Science and Information Systems Najran University, Najra 61441, Saudi Arabia*
[3]*Department of Computing and Technology, Abasyn University Peshawar, Peshawar 25000, Pakistan*
[4]*Secure Cyber Systems Research Group, WMG, University of Warwick, Coventry CV4 7AL, UK*
[5]*Information Systems Department, Faculty of Management Comenius University in Bratislava Odbojárov 10, Bratislava 82005 25, Slovakia*

Correspondence should be addressed to Muhammad Abul Hassan; abulhassan900@gmail.com

Received 27 April 2022; Accepted 20 June 2022; Published 13 July 2022

Academic Editor: Yousaf Bin Zikria

DeletedCoronary heart disease has an intense impact on human life. Medical history-based diagnosis of heart disease has been practiced but deemed unreliable. Machine learning algorithms are more reliable and efficient in classifying, e.g., with or without cardiac disease. Heart disease detection must be precise and accurate to prevent human loss. However, previous research studies have several shortcomings, for example,take enough time to compute while other techniques are quick but not accurate. This research study is conducted to address the existing problem and to construct an accurate machine learning model for predicting heart disease. Our model is evaluated based on five feature selection algorithms and performance assessment matrix such as accuracy, precision, recall, F1-score, MCC, and time complexity parameters. The proposed work has been tested on all of the dataset'sfeatures as well as a subset of them. The reduction of features has an impact on theperformance of classifiers in terms of the evaluation matrix and execution time. Experimental results of the support vector machine, K-nearest neighbor, and logistic regression are 97.5%,95 %, and 93% (accuracy) with reduced computation timesof 4.4, 7.3, and 8seconds respectively.

1. Introduction

Chronic heart diseases are one of the most dangerous and life-threatening worldwide. The fundamental cause of heart failure is narrowing and blockage of coronary arteries, where the heart fails to supply enough blood to other organs [1, 2]. The coronary arteries must be accessible to supply blood to the heart. According to a recent study, heart disease is the most common disease in the United States and worldwide with a high percentage of heart disease patients [3]. Common symptoms are shortness of breath, swelling feet, and tiredness [4]. Junk food with a maximum number of cholesterols, smoking, poor nutrition, high blood pressure, and physical inactivity increase the risks of heart disease [5]. Heartburn, stroke, and heart attack are all symptoms of coronary artery disease (CAD). Other heart disorders include heart rhythm problems, congenital heart disease, congestive heart failure, and cardiovascular disease. Traditional methods for detecting cardiac disease were used [6]. Lack of medical understanding and diagnostic instruments, on time detecting, and treating heart disease in poor countries is very difficult [7, 8]. The main motivation behind the research study is to propose a comprehensive and precise diagnosis technique for heart disease to avoid loss of lives. Cardiovascular disease is the leading cause of death in both developed and developing countries. According to the

WHO, 17.90 million people died from cardiovascular disease (CVD) in 2016, accounting for 30% of all deaths globally. Moreover, 0.2 million Pakistanis per year face death and death counts are still uplifting per year. According to the European Society of Cardiology (ESC), there are 26.5 million people in Europe who suffer from heart disease, with 3.8 million new cases being discovered each year. Heart disease kills 50–55% of patients in the first year, and treatment costs 4% of the yearly healthcare expenditure [9]. Invasive diagnostic procedures relied on a patient's medical history, physical examination results, and an examination of symptoms to make a diagnosis of heart disease [10]. Traditional methods like angiography are regarded as the most precise practice when it comes to detecting heart abnormalities but still facing certain limitations, such as high costs, various other side effects, and a high level of technical expertise is required, and most importantly it is much expensive, computationally difficult, and take time to assess [11, 12], to overcome the limitations of conventional invasive-based approaches for detecting cardiac disease. Predictive machine learning and deep learning algorithms were used to construct noninvasive Internet of Medical Thing (IoMT) [13–16], smart healthcare systems such as KNN, SVM, NB, DT, LR, RF, and ANN [17–22]. As a result, the death rate among individuals with heart disease has exponentially dropped per year.

The main objectives of this research study are as follows:

(i) To develop an intelligent medical decision system for the identification of cardiac illness on time.

(ii) Machine learning classification methods such as decision tree (DT), stochastic gradient descent (SGD), K-nearest neighbor (KNN), naive Bayes (NB), random forest (RF), logistics regression (LR), and support vector machine (SVM) are used to select the best model for early heart disease diagnosis.

(iii) Feature selection such as LASSO, ANOVA, MultiSURF, variance threshold, and mutual information to identify the most important and linked features that properly reflect the pattern of the desired target.

(iv) Cleveland hospital datasets related to heart disease are utilized.

The rest of the paper is organized as follows: Section 2 provides an overall literature review, materials and methods are explained in Section 3, results and discussion are discussed in Section 4, and Section 5 provides a conclusion.

2. Literature Review

Over time experts and practitioners have shown keen interest in diagnosing heart disease by employing classical machine learning techniques. Experts usually utilize a classification approach to create a heart disease diagnosis model in their research study [5, 23–38]. The machine learning model can diagnose heart failure with 99% accuracy, according to preliminary computational results as shown in Table 1.

Current research has imbalanced distribution, e.g., some approaches are accurate but required a long time for computation, and some techniques responded on time but are not very accurate to diagnose such serious disease. As a result, there is a great deal of work to improve the performance evaluation rate in this area.

3. Materials and Methods

The suggested approach aims to distinguish patients with or without cardiac disease. Both complete and selective features are enforced to investigate predictive models. Important features are identified using methods, e.g., LASSO, ANOVA, MultiSURF, variance threshold, and mutual information. K-nearest neighbor (KNN), support vector machine (SVM), decision tree (DT), random forest (RF), logistic regression (LR), stochastic gradient descent (SGD), and naive Bayes (NB) machine learning algorithms are deployed in the system for classification. Structure based on four steps, including exploratory data analysis, feature selection, ML classifiers, and performance evaluation matrix approach, is adopted. Algorithm 1 and Figure 1depict the proposed system's framework.

3.1. Preprocessing. Cleaning data is very important to achieve maximum accuracy and actual efficiency of machine learning algorithms. Different data preparation techniques are used to ensure each and every features must have the same coefficient. Moreover, standard scalar assures that each feature has the same mean, while min-max scalar shifts of data are set between 0 and 1, and lastly the row with missing values is erased.

3.2. Feature Selection. Precise and accurate feature selection is a very important parameter because it improves classification accuracy with minimum time complexity. LASSO, ANOVA, MultiSURF, variance threshold, and mutual information feature selection algorithms are used to select features from the dataset.

In the LASSO algorithm, some coefficients (feature) become zero, and are removed from the feature subset, derived from equations (1)–(6), while ANOVA compares the mean of two or more groups that are statistically distinct, derived from equations (7)–(11). MultiSURF is the most reliable feature selection algorithm explained in equations (12) and (13) and can be used for explicitly detecting pure 2-way interactions across a wide range of problems. Variance threshold is efficient in eliminating all features with variance below a certain threshold evaluated from equation (20). Lastly, we used mutual information in the feature selection phase to find dimensionless quantities with units of bits that measure "how much one random variable provides information about another." Mathematical modulation behind mutual information is explained in equation (15)–(20).

We have N number of samples $\{(x\square, y\square)\}\ \square\square_{=1}$ in the linear regression, where each $x\square = (x\square_1,\ldots,x\square p)$ is a p-dimensional vector of features, and each $y\square \in \mathbb{R}$ is the corresponding response variable. Our goal is to use a linear

Table 1: Previous literature review.

Reference	Heart disease type	Application	ML algorithm	Approach	Evaluations (%)	Data
[23]	Coronary disease	Classification	CA, BA	Undersampling	71.1	425 patients data
[24]	General heart disease	Classification	MLP	Undersampling	80	Cleveland dataset
[25]	General heart disease	Classification	ANN	Sampling	84	Cleveland dataset
[26]	General heart disease	Three-phase system for the prediction	ANN	Data sampling	85	Uci
[5]	Heart disease	Ensemble-based predictive model	ANN	Undersampling	91	Cleveland heart disease
[27]	Coronary Heart disease	Adaptive fuzzy ensemble	GA, MS-pso	Feature selection	92.31	Public dataset
[28]	Coronary artery disease	Classification	SVM, NB	Feature selection	96	Z-Alizadeh sani dataset
[29]	Cardiac disease	Classification	SVM, DT, KNN, etc.	Focal loss	86	Cleveland heart disease
[30]	Cardiac arrest	Scoring system classification	SVM	Undersampling	78.8	1386 records
[31]	Heart disease (general)	Detection	NB, SMO	Features selection	83	Cleveland dataset
[32]	Coronary heart disease	Predication	SVM, KNN, etc.	SMOTE	72	African heart disease data
[33]	Arrhythmia	Diagnosing	SVM, KNN, DT, RF	SMOTE	92	MIT-BIH
[34]	Heart arrhythmia	Detection	XGBoost classifier	Undersampling	87	Biobank UK dataset
[35]	Chronic heart failure (HF)	Incremental and boosting features value	DT, RF, SVM, KNN, LMT	Undersampling	89	487 patient data
[36]	Cardiovascular diseases	Classification	RF, DT	SMOTE	91	4270 patients data
[37]	Heart disease (general)	Features method	Lda, KNN, SVM, RF	Sampling	84	UCI dataset
[38]	Heart arrhythmia	Classification	Marine predators algorithm, SGD, CNN	Sampling	99.47	MIT-BIH arrhythmia, European, INCART
[39]	Heart arrhythmia	Classification	Marine predators algorithm, DNN, CNN	Sampling	99	MIT-BIH, EDB, and INCART
[38]	Heart arrhythmia	Classification	Manta ray foraging optimization, SVM, LBP, HOS	Sampling	98.26	MIT-BIH arrhythmia

mixture of features to approximate the response variable y□. Then the cost function (or loss function) must be optimized by using MSE as a cost function to determine the best fit line.

$$\textbf{LASSO} = \eta(x_i) = \beta_0 + \sum_{5=1}^{P} x_{ij}\beta_j, \tag{1}$$

$$\arg\min \beta 0, \beta \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \beta 0 - \sum_{j=1}^{p} x_{ij}\beta_j \right) \hat{2} \right\}. \tag{2}$$

The following equation shows the closed form solution that determines the coefficients of the aforesaid cost function. LASSO reduces the coefficients of redundant variables to zero, allowing the direct feature method. The LASSO cost function is as follows:

$$\beta = \left(X^T X\right)^{-1} X^T Y, \tag{3}$$

$$\arg\min \beta 0, \beta \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \beta 0 - \sum_{j=1}^{p} x_{ij}\beta_j \right) \hat{2} \right\} \lambda \sum_{j=1}^{p} \left|\beta_j\right|, \tag{4}$$

$$E(\beta) + \lambda R(\beta). \tag{5}$$

In equation (6) argmin finds values where the expression $E(\beta) + R(\beta)$ is minimum. The sparsity (β^*) of a model is defined by the number of parameters in β^* that are exactly equal to zero. In real-world problems, we need the model to take up only the most useful traits. LASSO regularization yields sparse solutions, which automatically choose features.

(1) **Input:** K, f
(2) **Output:** K Classification with matrix evaluation
(3) **if** $D_k \neq 0$ **then**
(4) **procedure** (features (f), K)
(5) f [V] nk+1/k ←Null
(6) z = (x - u)/s ←
(7) dataset f_r [0, 1] ← M
(8) end procedure
(9) **Procedure** Feature Extraction (f_k)
(10) (Lf_1, Lf_2, Lf_3. . .. . .Lf_n) ← L_k
(11) (Af_1, Af_2, Af_3. . .. . .Af_n) ← A_k
(12) (Mf_1, Mf_2, Mf_3. . .. . .Mfn) ← M_k
(13) (Vf_1, Vf_2, Vf_3. . .. . .. Vf_n) ← V_k
(14) (Mf_1, Mf_2, Mf_3. . .. . .. Mf_n) ← M_k
(15) **Return** (f_1, f_2, f_3. . .. . .f_n) ← k
(16) end procedure
(17) **procedure** C ((f_1, f_2, f_3. . .. . .f_n), G_k)
(18) M ⇐ T_k ((f_1, f_2, f_3. . .. . .f_n), G_k)
(19) P_k ⇐ TT_k (M, (f_1, f_2, f_3. . .. . .f_n))
(20) C_r ⇐ matrix (P_k, G_k)
(21) **return** P_k
(22) end procedure
(23) else
(24) $D_k = 0$ ← empty
(25) end if
(26) until: All the features (K) are Classified(C)
(27) Exit

Algorithm 1: Heart disease classification; take input features, preprocessing, feature selection, and classification.

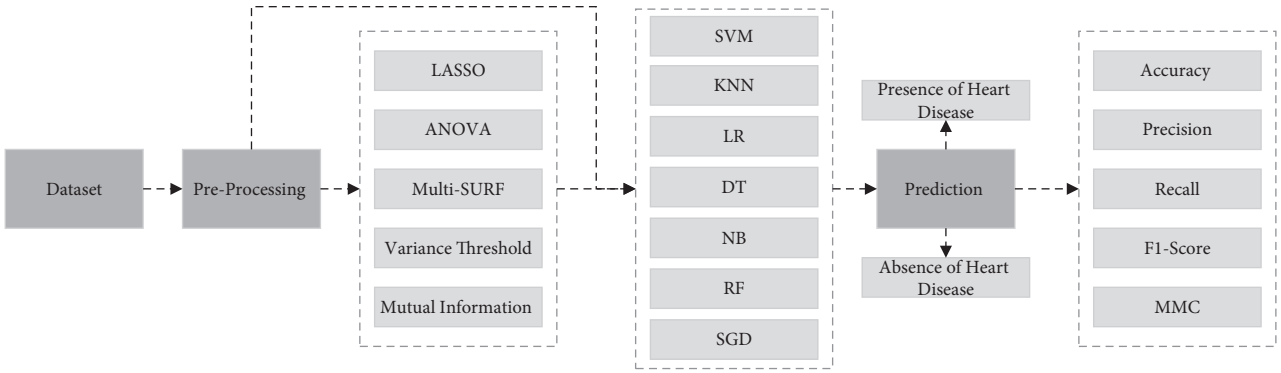

Figure 1: Proposed methodology to predict heart disease.

$$\beta^* = \arg\min [E(\beta) + \lambda R(\beta). \tag{6}$$

ANOVA makes use of the more traditional, standardized nomenclature. When we look at equations, we can see that the divisor has a degree of freedom (DF), the total is sum of squares (SS), we get mean square (MS), and the squared terms represent deviations from the sample mean. As a starting point, SS is partitioned into components that correspond to the model's effects.

$$\text{ANOVA} = \text{Ms}^2 = \frac{1}{n-1} \sum_i \left(y_i - \overline{y}\right)^2. \tag{7}$$

$$\text{ss}_{\text{Total}} = \text{ss}_{\text{Error}} + \text{ss}_{\text{Treatments}}. \tag{8}$$

Similarly, the number of degrees of freedom (DF) can be partitioned: one of these components specifies chi-squared distribution for error that represents the related sum of squares, and the same "treatments" have no effect if there is no value.

$$\mathrm{DF} = \mathrm{DF}_{\mathrm{Error}} + \mathrm{DF}_{\mathrm{Treatments}}. \tag{9}$$

In lieu of the more traditional one-way analysis of ANOVA, the following form can be used to express each piece of information.

$$y_{\mathrm{ij}} = \mu + T_j + \varepsilon_{\mathrm{ij}}, \tag{10}$$

$$\sum_{J=1}^{C} T_j = 0. \tag{11}$$

In the case of the Multi-SURF algorithm, each feature in the dataset is assigned to one of two groups. Inside the data collection, each feature should be scaled 0–1 and repeat the process m times with a p-long weight vector (W) of zeros. Then the feature vector (X) of a random instance and the feature vectors of the instances closest to X by Euclidean distance. It refers to the closest same-class instance, whereas it refers to the nearest different-class instance. In equation (13) we compute a two-tailed p-value using the cumulative distribution function to determine the number of cases that are close or distant.

$$\mathrm{MultiSURF} = w_i = w_i - \left(x_i - \mathrm{nearHit}_i\right)^2 + \left(x_i - \mathrm{nearMiss}_i\right)^2, \tag{12}$$

$$2\left(1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(1/2)} \mathrm{e} \frac{-x}{2} \mathrm{d}x\right) \approx 0.60. \tag{13}$$

The information-theoretic formula is used by the variance threshold algorithm to reduce dataset features. For a given feature subset Q, there are a variety of truth value assignments. A feature set Q divides training data into groups of instances with the same truth value into a set of training data instances. The entropy of positive P_i and negative n_i class values are calculated by using the below equation.

$$\mathrm{VarianceThreshold}(Q) = -\sum_{1=0}^{2|Q|-1} \frac{P_i + n_i}{|\mathrm{Sample}|} \left[\frac{P_i}{P_i + N_i} \log_2 \frac{P_i}{P_i + N_i} + \frac{n_i}{P_i + n_i} \log_2 \frac{n_i}{P_i + n_i}\right]. \tag{14}$$

Mutual information, as opposed to correlation coefficients, includes information on all linear and nonlinear dependencies. However, if the joint distribution of X and Y is bivariate normal and both marginal distributions are normally distributed, the relationship between I and p is precise.

$$\text{Mutual Information} = H\left(X_i\right) = \frac{1}{2} \log\left[2\pi e \sigma_1^2\right), \tag{15}$$

$$\frac{1}{2} + \frac{1}{2} \log(2\pi) + \log(\sigma_1) i \epsilon \{1, 2\}, \tag{16}$$

$$H\left(x_1 x_2\right) = \frac{1}{2} \log(2\pi e)^2 |\Sigma|, \tag{17}$$

$$1 + \log(2\pi) + \log\left(\sigma_1 \sigma_2\right) + \frac{1}{2} \log\left(1 - p^2\right), \tag{18}$$

$$I\left(X_1, X_2\right) = H\left(X_1\right) + \left(X_2\right) - H\left(X_1, X_2\right), \tag{19}$$

$$I = -\frac{1}{2} \mathrm{Log}(1 - P)^2. \tag{20}$$

3.3. Classification. Heart patients and healthy patients are separated into groups using machine learning classification methods. In this phase, we will take a look at a few prominent classification approaches as well as the theoretical basis of those methods.

3.3.1. Support Vector Machine (SVM). SVM is an ML classification technique; this has mainly been used to solve classification issues. It uses a maximum margin strategy to solve a complex quadratic problem, and is employed in a variety of applications due to its high classification performance. Moreover, SVM is best suited for identifying the best hyperplane to separate the data, as shown in equations (21)–(23).

$$w.x + b = 0, \tag{21}$$

$$X_1 * w + b \leq 1, \quad \mathrm{yi} = -1. \tag{22}$$

$$X_1 * w + b \geq 1, \quad \mathrm{yi} = +1. \tag{23}$$

3.3.2. Naïve Bayes (NB). The NB method uses the conditional probability theorem as can be seen in equation (24), to classify new feature vectors and also find their conditional probability values. The conditionality likelihood of each vector is used to calculate the new vector class and is usually utilized for text-related problem classification.

$$p(x|y) = \frac{p(y \mid x) p(x)}{P(y)}. \tag{24}$$

3.3.3. Decision Tree (DT). DT is also an ML approach where each node is a leaf node with internal and external nodes connected. The internal nodes make decisions and send child nodes to the next node, whereas the leaf node has no child nodes, and is labeled derived from the following equations:

$$I = -\sum_{C} P(C) \log_2 p(c), \tag{25}$$

$$\mathrm{Gain}(A) = I - I(\mathrm{res}), \tag{26}$$

$$I = 1 - \sum_{j} p(c)^2, \tag{27}$$

$$G = 1 - \sum \frac{cj}{c} I(c), \tag{28}$$

$$d = \sum_{i=1}^{k} |(x - yi)^2. \tag{29}$$

3.3.4. K-Nearest Neighbor (KNN). KNN uses the similarity of new input to the incoming input samples in the training set and to predict a new input's class label, as shown in the following equation:

$$d = \sum_{i=1}^{k} |(x - yi)^2. \tag{30}$$

3.3.5. Logistic Regression (LR). Binary classification problems are solved using a logistic regression technique, which predicts values for variables 0 and 1, and classifies them into two groups: negative (0) or positive (1). A threshold value of 0.5 is used in the multi-classification approach to predict decimal numbers, which is then used to classify the two classes, e.g., 0 and 1. Hypothesis if threshold ≥0.5 predicts 1, indicating that the patient has heart disease (cardiomyopathy). The mathematical representation of logistic regression is explained in the following equations:

$$P(x) = \frac{1}{1 + e^{-((x-\mu)/s)}}, \tag{31}$$

$$P(x) = \frac{1}{1 + e^{-(\beta 0 - \beta 1 x)}}, \tag{32}$$

$$-\text{yk In } pk - (1 - \text{yk}) \text{In}(1 - pk), \tag{33}$$

$$e = \sum_{k:\ \text{yk}=1} \text{In}(\text{pk}) + \sum_{k:\ \text{yk}=0} \text{In}(1 - \text{pk}), \tag{34}$$

$$\sum_{k=1}^{k} (\text{yk} \text{In}(\text{pk}) + (1 - \text{yk}) \text{In}(1 - \text{pk})), \tag{35}$$

$$L = \prod_{k:\ yk=1} (\text{pk}) \prod_{k:\ yk=0} (1 - \text{pk}). \tag{36}$$

3.3.6. Random Forest. A random forest is a meta estimator explained in equations (37) and (38), that uses averaging to improve prediction accuracy while minimizing overfitting. The subsample size is determined by the max-samples option, and each tree uses the entire dataset.

$$\widehat{f} = \frac{1}{B} \sum_{B=1}^{B} \text{fb}(x'), \tag{37}$$

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B} \left(\text{fb}(x') \widehat{f}\right)^2}{B - 1}}. \tag{38}$$

3.3.7. Stochastic Gradient Descent (SGD). SGD has received significant attention, despite its long history in machine learning applications. Convex loss faced in SVM and LR is addressed by SGD. This technique (SGD) provides a quick and easy technique to fit linear classifiers and regressions in the context of large-scale learning. Equations (39)–(41) explain the SGD technique to provide a quick and best-fit machine learning classifier.

$$w = w - \eta \nabla Q(w), \tag{39}$$

$$w - \frac{\eta}{n} \sum_{i=1}^{n} \nabla Q_i(w), \tag{40}$$

$$W = w - \eta \nabla Q_i(w). \tag{41}$$

3.3.8. Performance Matrix. Several performance matrices are explained in equations (42)–(46), including accuracy, recall, precision, F1-score, and Matthews correlation coefficient (MCC). These evaluation parameters are used to check the performance of our proposed approach with other algorithms.

$$\text{Accuracy} = \frac{\text{Tp} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100, \tag{42}$$

$$\text{Precision} = \frac{\text{Tp}}{\text{TP} + \text{FP}} * 100, \tag{43}$$

$$\text{Recall} = \frac{\text{Tp}}{\text{TP} + \text{FN}} * 100, \tag{44}$$

$$F1 - \text{score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{45}$$

$$\text{MCC} = \frac{\text{Tp} * \text{TN} - \text{FP} * \text{FN}}{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})} * 100, \tag{46}$$

$$\text{Computational Time} \quad \&9; = \text{software Time} * \text{Number of Features} * \text{Clock rate}. \tag{47}$$

4. Result and Discussion

This section of the study provides various classification models and their statistical analysis. In the first phase, we compare the performance of LR, KNN, SGD, RF SVM, NB, and DT on the Cleveland heart disease dataset. In the second phase, we have employed LASSO, ANOVA, MultiSURF, variance threshold, and mutual information to pick relevant features. To evaluate the performance classifiers, all features were normalized and standardized before being supplied to classifiers.

Table 2: Classifier performance before feature selection.

Classifier		Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	MCC (%)	Time complexity (sec)
	1	75	80	73	76	51	
SVM	0	75	71	78	74	55	10.4
	Overall	75	75.5	75.5	75	53	
	1	67	64	69	66	35	
KNN	0	67	71	66	68	47	16.7
	Overall	67	67.6	67.5	67	41	
	1	71	76	69	72	42	
LR	0	71	63	73	69	33	12.2
	Overall	71	69.5	71	70.5	37.5	
	1	61	56	62	58	22	
DT	0	61	66	60	62	37	19.9
	Overall	61	61	61	60	29.5	
	1	70	75	68	71	41	
NB	0	70	66	72	69	39	24.7
	Overall	70	70.5	70	70	40	
	1	65	68	64	66	30	
RF	0	65	62	65	63	27	17.1
	Overall	65	65	64.5	64.5	28.5	
	1	69	76	66	71	39	
SGD	0	69	62	72	66	44	14.4
	Overall	69	69	69	68.5	41.5	

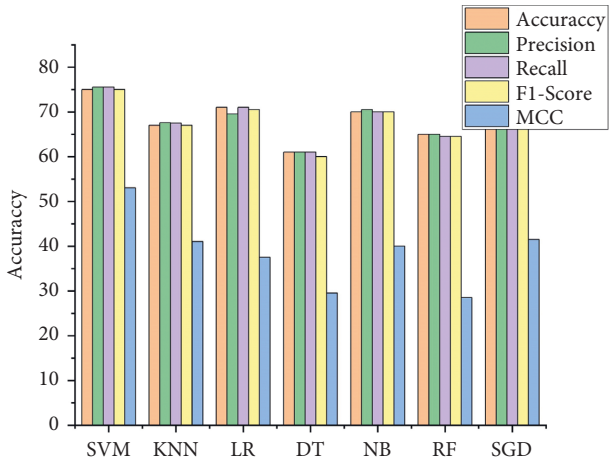

Figure 2: Result of the classifier with full feature.

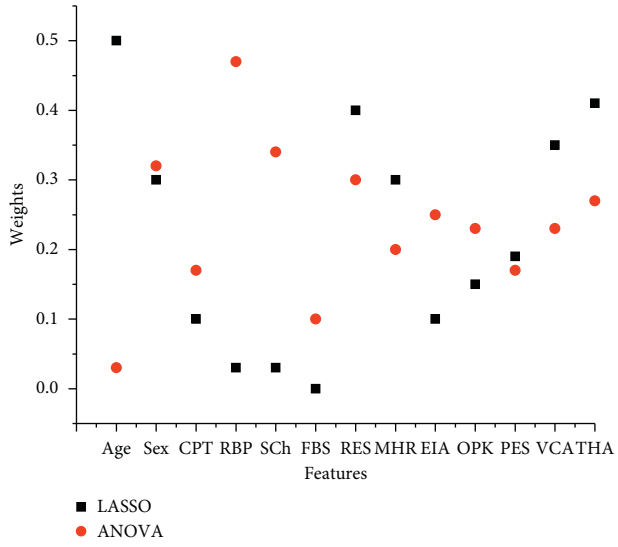

Figure 3: Feature selected with LASSO and ANOVA.

The features of the entire dataset were tested on selected machine learning classifiers in this experiment, where 7 : 3 ratio data is allocated for training (70%) and testing (30%).

In Table 2 and Figure 2, the SVM shows a good performance with 75% accuracy, 75.5% precision, 75.5% recall, 75% F1-score, 53% MMC, and 10.4 seconds time complexity. Different K values are tested for the KNN classifier, and the best performance among all round is; 67% accuracy, 67.6% precision, 67.5% recall, F1-score 67%, MCC 41%, and time complexity of 16.7 second. The LR classifier achieved 71% accuracy, 69.5% precision, 71% recall, 70.5% F1-score, MCC 37.5%, and time complexity is 12.2 second. The DT classifier achieved 61% accuracy, 61% precision, 61% recall, 60% F1-score, MCC 29.5%, and time complexity is 19.9 second. The NB classifier achieved 70% accuracy, 70.5% precision, 70% recall, 70% F1-score, MCC 40%, and time complexity is 24.7 second. The RF classifier achieved 65% accuracy, 65% precision, 64.5% recall, 64.5% F1-score, MCC 28.5%, and time complexity is 17.1 second. The SGD classifier achieved 69% accuracy, 69% precision, 69% recall, 68.5% F1-score, MCC 41.5%, and time complexity is 14.4 second.

Based on their weight, LASSO and ANOVA select different features from the complete dataset. LASSO is used to select the five most important features namely SEX, RES, MHR, VCA, and THA. ANOVA select features, e.g., SEX, RBP, SCH, RES, and THA, as can be seen in Table 3 and Figure 3. We analyzed classifiers on a variety of chosen features and performances are very efficient.

The five most relevant features are selected and to be utilized in the second group of feature selection, namely MultiSURF, variance threshold, and mutual information, as

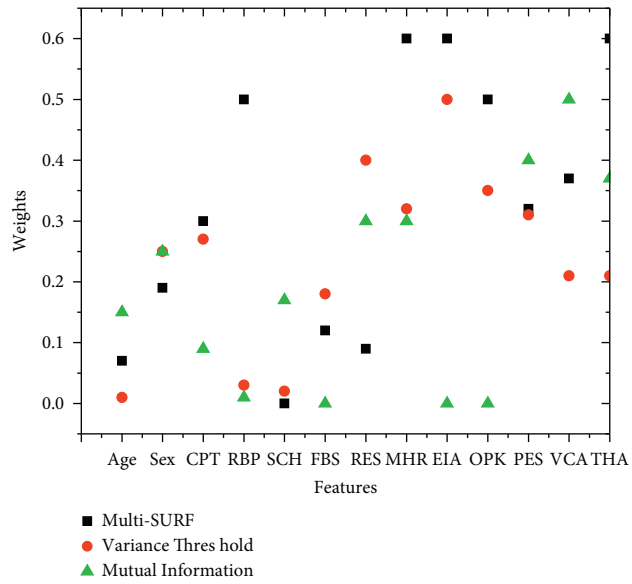

Figure 4: Selected feature with Multi-SURF, variance threshold, and mutual information.

Table 3: Selected feature rank.

Number	Algorithm		Feature name	Feature code	Rank
		1	Gender	SEX	0.3
		2	Resting electrocardiography	RES	0.4
1	LASSO	3	Maximum heart rate	MHR	0.3
		4	Number of major vessels	VCA	0.35
		5	Thallium scan	THA	0.41
		1	Gender	SEX	0.32
		2	Level of BP	RBP	0.47
2	ANOVA	3	Serum cholesterol	SCH	0.34
		4	Resting electrocardiography	RES	0.3
		5	Thallium scan	THA	0.27
		1	Level of BP	RBP	0.5
		2	Maximum heart rate	MHR	0.6
	Multi SURF	3	Exercise-induced angina	EIA	0.6
		4	Old peak	OPK	0.5
		5	Thallium scan	THA	0.6
		1	Resting electrocardiography	RES	0.4
		2	Maximum heart rate	MHR	0.32
4	Variance threshold	3	Exercise-induced angina	EIA	0.5
		4	Old peak	OPK	0.35
		5	Slope of the peak exercise	PES	0.31
		1	Resting electrocardiography	RES	0.3
		2	Maximum heart rate	MHR	0.3
5	Mutual information	3	Slope of the peak exercise	PES	0.4
		4	Number of major vessels	VCA	0.5
		5	Thallium scan	THA	0.37

shown in Table 3 and Figure 4. Multi-SURF selects RBP, MHR, EIA, OPK, and THA features from the dataset. RES, MHR, EIA, OPK, and PES features are the most prominent features for variance threshold. Moreover, RES, MHR, PES, VCA, and THA are chosen by mutual information select features which is the final and most essential feature selection algorithm.

As demonstrated in Figures 3 and 4, after features selection, the five most important features are tested on different machine learning classifiers, with a 7 : 3 ratio set for the training (70%) and testing (30%). In Table 4 and Figure 5, SVM shows a good performance by using a confusion matrix with 97.5% accuracy, 97% precision, 97% recall, 97% F1-score, 95% MMC, and 4.4 seconds time complexity. Different K values are applied for the KNN classifier and best among them are 95% accuracy, 95% precision, 95% recall, F1-score 95%, 88.5% MCC, and 7.3 seconds time complexity. The LR classifier achieved 93% accuracy, 93.5% precision,

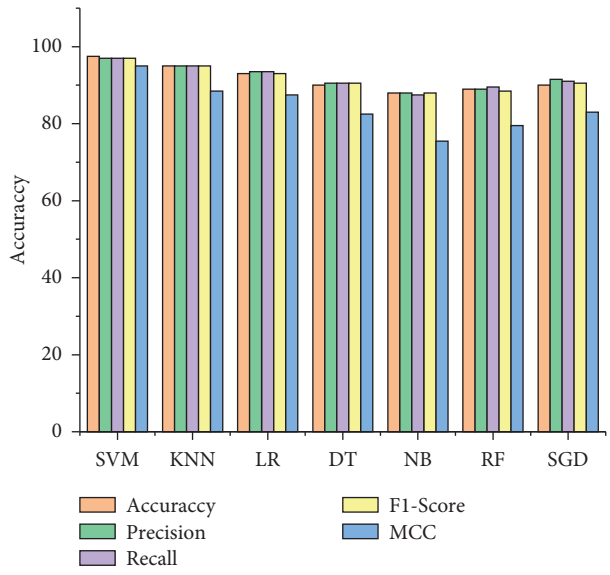

Figure 5: Result of the classifier with selected features.

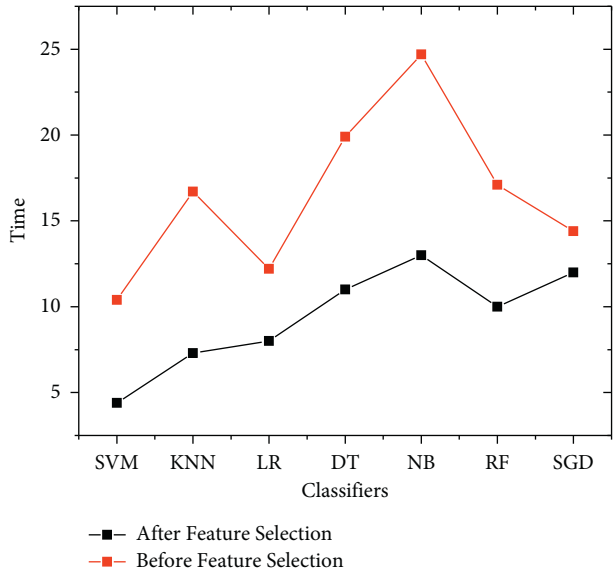

Figure 6: Time complexity with full features and selected features.

93.5% recall, 93% F1-score, 87.5% MCC, and 8 seconds time complexity. The DT classifier has achieved 90% accuracy, 90.5% precision, 90.5% recall, 90.5% F1-score, 82.5% MCC, and 11 seconds time complexity. The NB classifier achieved 88% accuracy, 88% precision, 87.5% recall, 88% F1-score, 75.5% MCC, and 13.9 seconds time complexity. The RF classifier achieved 89% accuracy, 89% precision, 89.5% recall, 88.5% F1-score, 79.5% MCC, and 10 seconds time complexity. The SGD classifier achieved 90% accuracy, 91.5% precision, 91% recall, 90.5% F1-score, 83% MCC, and 12 seconds time complexity.

Figure 6 depicts the classifier parameters for overall features and five main characteristics to demonstrate time complexity of each classifier. The SVM algorithm has 4.4 seconds for selected features and 10.4 seconds for all other features in the dataset. KNN has 7.3 and 16.7 seconds, respectively. The LR algorithm has 8 and 12.2 seconds with and without features, the DT algorithm has 11 and 19.9 seconds, and the NB algorithm has 13.9 and 24.7 seconds. RF processing time for classifying the dataset is 10 and 17.1 seconds, and lastly, SGD has 12 and 14.4 seconds, respectively.

Table 5 illustrates an increase in SVM classification accuracy from 75% to 97.5% on minimized features. Similarly, the accuracy of KNN improved from 67% to 95% with reduced features, LR increased from 71% to 91%, DT increased from 61% to 90%, NB increased from 70% to 88%, RF increased from 65% to 89%, and SGD increased from 69% to 90%. As a result, the feature selection algorithms select significant features that boost the performance of the

Table 4: Selected feature result.

Classifier		Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	MCC (%)	Time complexity (sec)
	1	97	96	98	97	94	
SVM	0	98	98	96	97	96	4.4
	Overall	97.5	97	97	97	95	
	1	95	92	98	95	90	
KNN	0	95	98	92	95	87	7.3
	Overall	95	95	95	95	88.5	
	1	93	93	94	93	87	
LR	0	93	94	93	93	88	8
	Overall	93	93.5	93.5	93	87.5	
	1	90	90	93	91	81	
DT	0	90	91	88	90	84	11
	Overall	90	90.5	90.5	90.5	82.5	
	1	88	90	86	88	76	
NB	0	88	86	89	88	75	13.9
	Overall	88	88	87.5	88	75.5	
	1	89	92	87	89	78	
RF	0	89	86	92	88	81	10
	Overall	89	89	89.5	88.5	79.5	
	1	90	85	95	90	82	
SGD	0	90	96	87	91	84	12
	Overall	90	91.5	91	90.5	83	

Table 5: Improved accuracy result.

Classifier	Before feature selection (%)	After feature selection (%)
SVM	75	97.5
KNN	67	95
LR	71	93
DT	61	90
NB	70	88
RF	65	89
SGD	69	90

Table 6: Comparative analysis.

Matric Classifier	Proposed work Accuracy	Previous work Accuracy
SVM	97.5%	88%
KNN	95%	81%
LR	91%	89%
DT	90%	83%
NB	88%	83%
RF	89%	66%
SGD	90%	N/A

classifier and reduce execution time to effectively diagnose heart disease prediction.

4.1. Comparative Analysis. We employed several feature selection and machine learning approaches in the classification phase. The results demonstrated that our suggested methods produce efficient outcomes in terms of all performance matrices with minimum computational time. In the end, based on statistical data, we conclude that our proposed approach has improved the overall performance of algorithms as can be seen in Table 6.

5. Conclusion

This research study proposed a machine-learning-based cardiac disease classification system. Decision tree (DT), stochastic gradient descent (SGD), K-nearest neighbor (KNN), naive Bayes (NB), random forest (RF), logistics regression (LR), and support vector machine (SVM) were used to classify the Cleveland heart disease dataset collected from Cleveland hospitals. The novelty of this proposed work is the development of a diagnosis system for heart disease patients. Feature selection algorithms such as LASSO, ANOVA, MultiSURF, variance threshold, and mutual information are utilized before supplying data for the training and test phase, main motivation behind this approach is to improve the response time of each algorithm. Performance evaluation matrices, e.g., accuracy, precision, recall, F1-score, and MMC, were used to compare the different classifier performances. In addition, the proposed approach is evaluated on a 5-feature algorithm with 7 classifiers and 5 performance evaluation metrics and have shown efficient performance (refer to section 4). A machine learning classification model is used in this study. SVM, KNN, and LR models all perform well with specific features and can improve classification accuracy while also reducing the overall processing time. The findings are consistent with earlier research. In the future, we will apply federated learning and blockchain algorithms to generate an effective and efficient diagnosing system.

Data Availability

The data used to support the findings of the study are included in the article https://www.kaggle.com/datasets/aavigan/cleveland-clinic-heart-disease-dataset.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research work was supported by the Information Systems Department, Faculty of Management Comenius University in Bratislava Odbojárov 10, 82005 Bratislava 25, Slovakia.

References

[1] B. Ziaeian and G. C. Fonarow, "Epidemiology and aetiology of heart failure," *Nature Reviews Cardiology*, vol. 13, no. 6, pp. 368–378, 2016.

[2] K. Polat and S. Güneş, "Artificial immune recognition system with fuzzy resource allocation mechanism classifier, principal component analysis and FFT method based new hybrid automated identification system for classification of EEG signals," *Expert Systems with Applications*, vol. 34, no. 3, pp. 2039–2048, 2008.

[3] P. A. Heidenreich, J. G. Trogdon, O. A. Khavjou et al., "Forecasting the future of cardiovascular disease in the United States," *Circulation*, vol. 123, no. 8, pp. 933–944, 2011.

[4] T. Kohn, S. Shabtaie, and R. Ramasamy, "Effect of strict sperm morphology on intrauterine insemination pregnancy success: a systematic review and meta-analysis," *Fertility and Sterility*, vol. 106, no. 3, pp. e91–e92, 2016.

[5] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," *Expert Systems with Applications*, vol. 36, no. 4, pp. 7675–7680, 2009.

[6] L. A. Allen, L. W. Stevenson, K. L. Grady et al., "Decision making in advanced heart failure," *Circulation*, vol. 125, no. 15, pp. 1928–1952, 2012.

[7] H. Yang and J. Garibaldi, "A hybrid model for automatic identification of risk factors for heart disease," *Journal of Biomedical Informatics*, vol. 58, pp. S171–S182, 2015.

[8] R. Alizadehsani, J. Habibi, Z. A. Sani et al., "Diagnosis of coronary artery disease using data mining based on lab data and echo features," *Journal of Medical and Bioengineering*, vol. 1, no. 1, pp. 26–29, 2012.

[9] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm," *Computer Methods and Programs in Biomedicine*, vol. 141, pp. 19–26, 2017.

[10] O. Samuel, G. Asogbon, A. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy_ AHP for heart failure risk prediction," *Expert Systems with Applications*, vol. 68, pp. 163–172, 2017.

[11] K. Phegade, "Heart attack prediction system using artificial neural network," *International Journal for Research in Applied Science and Engineering Technology*, vol. 7, no. 2, pp. 971–974, 2019.

[12] C. da Costa, G. da Costa Linch, and E. Nogueira de Souza, "Nursing diagnosis based on signs and symptoms of patients with heart disease," *International Journal of Nursing Knowledge*, vol. 27, no. 4, pp. 210–214, 2016.

[13] M. K. Hasan, T. M. Ghazal, A. Alkhalifah et al., "Fischer linear discrimination and quadratic discrimination analysis–based data mining technique for Internet of things framework for healthcare," *Frontiers in Public Health*, vol. 9, 2021.

[14] T. M. Ghazal, "Hep-pred: hepatitis C staging prediction using fine Gaussian SVM," *Computers, Materials & Continua*, vol. 69, no. 1, pp. 191–203, Article ID 015436, 2021.

[15] M. K. Hasan, S. Islam, I. Memon et al., "A novel resource oriented DMA framework for Internet of medical things devices in 5G network," *IEEE Transactions on Industrial Informatics*, p. 1, 2022.

[16] S. Y. Siddiqui, A. Haider, T. M. Ghazal et al., "IoMT cloud-based intelligent prediction of breast cancer stages empowered with deep learning," *IEEE Access*, vol. 9, pp. 146478–146491, Article ID 3123472, 2021.

[17] D. Hales and B. Edmonds, "Applying a socially inspired technique (tags) to improve cooperation in P2P networks," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 35, no. 3, pp. 385–395, 2005.

[18] A. Methaila, P. Kansal, H. Arya, and P. Kumar, "Early Heart Disease Prediction Using Data Mining Techniques," in *Proceedings of the Computer Science & Information Technology (CS & IT)*, August 2014.

[19] M. Rizwan, A. Shabbir, A. R. Javed et al., "Risk monitoring strategy for confidentiality of healthcare information," *Computers & Electrical Engineering*, vol. 100, Article ID 107833, 2022.

[20] F. Sajid, M. A. Hassan, A. A. Khan et al., "Secure and Efficient Data Storage Operations by Using Intelligent Classification Technique and RSA Algorithm in IoT-Based Cloud Computing," *Scientific Programming*, vol. 2022, Article ID 2195646, 2022.

[21] M. Rizwan, A. Shabbir, A. R. Javed et al., "Risk monitoring strategy for confidentiality of healthcare information," *Computers and Electrical Engineering*, vol. 100, Article ID 107833, 2022.

[22] F. Sajid, M. A. Hassan, A. A. Khan et al., "Secure and efficient data storage operations by using intelligent classification technique and RSA algorithm in IoT-based cloud computing," *Scientific Programming*, Article ID 2195646, 10 pages, 2022.

[23] F. Al Badarin and S. Malhotra, "Diagnosis and prognosis of coronary artery disease with SPECT and PET," *Current Cardiology Reports*, vol. 21, no. 7, 2019.

[24] S. Wankhade and P. Shahabade, "Hiding secret data through steganography IN VOIP," *International Journal of Computer and Communication Technology*, vol. 8, no. 2, pp. 116–120, 2017.

[25] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Systems with Applications*, vol. 35, no. 1-2, pp. 82–89, 2008.

[26] E. Olaniyi, O. Oyedotun, and K. Adnan, "Heart diseases diagnosis using neural networks arbitration," *International Journal of Intelligent Systems and Applications*, vol. 7, no. 12, pp. 75–82, 2015.

[27] A. Paul, P. Shill, M. Rabin, and K. Murase, "Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease," *Applied Intelligence*, vol. 48, no. 7, pp. 1739–1756, 2017.

[28] R Alizadehsani, M. J. Hosseini, A. Khosravi et al., "Non-invasive detection of coronary artery disease in high-risk patients based on the stenosis prediction of separate coronary arteries," *Computer Methods and Programs in Biomedicine*, vol. 162, pp. 119–127, 2018.

[29] A. Haq, J. Li, M. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2018, pp. 1–21, Article ID 3860146, 2018.

[30] N. Liu, Z. Lin, J. Cao et al., "An intelligent scoring system and its application to cardiac arrest prediction," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1324–1331, 2012.

[31] J. Nahar, T. Imam, K. Tickle, and Y. Chen, "Computational intelligence for heart disease diagnosis: a medical knowledge driven approach," *Expert Systems with Applications*, vol. 40, no. 1, pp. 96–104, 2013.

[32] H. Khdair and N. Dasari, "Exploring machine learning techniques for coronary heart disease prediction," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, 2021.

[33] S. Ketu and P. Mishra, "Empirical analysis of machine learning algorithms on imbalance electrocardiogram based arrhythmia dataset for heart disease detection," *Arabian Journal for Science and Engineering*, vol. 47, no. 2, pp. 1447–1469, 2021.

[34] M. Rezaei, J. Woodward, J. Ramírez, and P. Munroe, "A novel two-stage heart arrhythmia ensemble classifier," *Computers*, vol. 10, no. 5, p. 60, 2021.

[35] D. K. Plati, E. E. Tripoliti, A. Bechlioulis et al., "A machine learning approach for chronic heart failure diagnosis," *Diagnostics*, vol. 11, no. 10, p. 1863, 2021.

[36] J. Minou, J. Mantas, F. Malamateniou, and D. Kaitelidou, "Classification techniques for cardio- vascular diseases using supervised machine learning," *Medical Archives*, vol. 74, no. 1, p. 39, 2020.

[37] K. Polat, "Similarity-based attribute weighting methods via clustering algorithms in the classification of imbalanced medical datasets," *Neural Computing & Applications*, vol. 30, no. 3, pp. 987–1013, 2018.

[38] E. Houssein, M. Hassaballah, I. Ibrahim, D. AbdElminaam, and Y. Wazery, "An automatic arrhythmia classification model based on improved Marine Predators Algorithm and Convolutions Neural Networks," *Expert Systems with Applications*, vol. 187, Article ID 115936, 2022.

[39] E. Houssein, D. Abdelminaam, I. Ibrahim, M. Hassaballah, and Y. Wazery, "A hybrid heartbeats classification approach based on marine predators algorithm and convolution neural networks," *IEEE Access*, vol. 9, pp. 86194–86206, 2021.

[40] E. Houssein, I. Ibrahim, N. Neggaz, M. Hassaballah, and Y. Wazery, "An efficient ECG arrhythmia classification method based on Manta ray foraging optimization," *Expert Systems with Applications*, vol. 181, Article ID 115131, 2021.