Iraqi Journal for Computer Science and Mathematics

Journal Homepage: http://journal.esj.edu.iq/index.php/IJCM
e-ISSN: 2788-7421 p-ISSN: 2958-0544

Heart disease classification using optimized Machine learning algorithms

Mohammad Abood Kadhim*1, Abdulkareem Merhej Radhi2

[1]Department of Computer Science, Al-Nahrain University, Baghdad, Iraq
[2]Computer Department/College of Science University AL-Nahrain, Baghdad, 10001, Iraq

*Corresponding Author: Mohammad Abood Kadhim

DOI: https://doi.org/10.52866/ijcsm.2023.02.02.004
Received November 2022; Accepted Janury 2023; Available online Februry 2023

ABSTRACT: Early detection of heart disease is exceptionally critical to saving the lives of human beings. Heart attack is one of the primary causes of high death rates throughout the world, due to the lack of human and logistical resources in addition to the high costs of diagnosing heart diseases which plays a key role in the healthcare sector, this model is suggested. In the field of cardiology, patient data plays an essential role in the healthcare system. This paper presents a proposed model that aims to identify the optimal machine learning algorithm that can predict heart attacks with high accuracy in the early stages. The concepts of machine learning are used for training and testing the model based on the patient's data for effective decision-making. The proposed model consists of three stages, the first stage is patient data collection and processing, and the second stage is data training and testing using machine learning algorithms Random Forest, Support Vector Machines, K-Nearest Neighbor, and Decision Tree) that show The best classification (94.958 percent) with the Random Forest algorithm and the third stage is optimized the classification results using one of the hyperparameters optimization techniques random search that shows The best accuracy was (95.4 percent) obtained also with RF.

Keywords: Machine learning (ML), K-Nearest Neighbor (KNN), Support Vector Machines (SVM), Decision tree (DT), Random Forest, hyperparameters.

1. INTRODUCTION

Health data needs to change the data looking for conduct and can be watched around the globe. Challenges confronted by numerous individuals are looking online for health data concerning illnesses, diagnosing diverse medicines that will take part of the time, and squandering money [1].

"World Health Organization" regards cardiovascular infections as the first reason for death universally. Up to 17.9 million individuals deceased in 2016, 31 percent of all deaths worldwide were due to coronary heart disease. Cardiovascular diseases (CVD) are heart and blood vessel problems. Heart attacks and strokes account for four out of every five deaths related to cardiovascular disease. People who are at risk for cardiovascular disease may have high blood pressure or be overweight or obese [2]. The heart, being one of the largest and most essential organs in the human body, needs extra care. Because the majority of diseases are connected to the heart, it is vital to predict cardiac disorders, which needs comparative research in this sector. Because most patients are dying due to the discovery of their diseases at an advanced stage owing to instrument inaccuracy, then more efficient prediction disease algorithms are needed [3]. Current heart disease diagnosis approaches are inefficient in early detection for a variety of reasons, including accuracy and execution time, thus researchers are aiming to create an effective methodology for the early identification of a cardiovascular illness. It is incredibly difficult to diagnose and treat heart disease when contemporary technology and medical specialists are unavailable [4]. In the research community, machine learning technologies have sparked a lot of interest. Machine learning approaches, as demonstrated in several recent papers, have the potential to provide high classification accuracy when compared to traditional data categorization procedures. Accurate prediction is critical since it can lead to adequate protection. The accuracy of predictions may fluctuate depending on the learning approach used. As a result, it's crucial to spot devices that can predict cardiac disease with great accuracy. The accuracy of the prediction achieved in this study is better as compared to the earlier research. Machine learning classification is one of the most practical approaches for generating assessments in both real-world and research contexts. Persistence is also used to evaluate the performance of various machine-learning approaches for the categorization of patients with

and without cardiac disease. In addition, the effectiveness of these techniques has been assessed using several categorization performance metrics [5].

This paper presents a proposed model that aims to design and implement an automated model to predict heart disease with high accuracy in the early stages. machine learning model with Hyper Parameter Optimization (HPO) Randomized Search technique was presented. To this end, the researchers create a pipeline of prediction algorithms for the clinical diagnosis of heart disease using machine learning technologies. To determine the characteristics of machine learning approaches, an experiment was undertaken. The heart disease datasets were obtained from the IEEE-data port data source. This dataset was chosen because it was curated by integrating five well-known cardiac disease datasets (Long Shoreline VA, Hungarian, Cleveland, Starlog, and Switzerland) and no research has worked on the same data previously. Four algorithms were used to generate prediction models for this experiment utilizing the provided dataset (Support Vector Machines, K-Nearest Neighbor, Decision tree, and Random Forest). Furthermore, the maximum accuracy is evaluated by the best approach discovered in this study to the highest accuracy obtained in previous research.

The remainder of the paper is organized as follows. In section 2 an overview of the related work is presented. Section 3 discusses the methodology. Research results and discussions are presented in Sections 4 and 5. Finally, the conclusion is given in Section 6.

2. RELATED WORKS

Several studies and experiments have been conducted on heart disease datasets. Below is a set of previous studies that show the dataset that researchers have worked on and which have been combined by identifying common characteristics. This combined data set will be worked on in this research.

In [6], Heart disease is a serious disease that is a leading cause of death in all countries of the world. However, it is difficult for doctors to predict such Diseases because they are complex and also considered expensive. In this research, the researchers proposed a clinical support system as an aid to medical specialists to predict and diagnose heart diseases and make the best decisions. Some Ml algorithms were applied in this study such as Naïve Bayes and KNN, SVM, RF, and DT to predict heart failure disease using risk factor data retrieved from medical files. Several experiments have been performed to predict the use of the HD UCI data set, and the best result with NB when using both cross-validations with an accuracy of 82.17 percent and split-test training and with an accuracy of 84.28 percent.

In [7], the goal of this study was to examine machine learning algorithms using several performance criteria to enhance accuracy. In the pre-processing stage, they use the mean value to replace missing values. The findings show that using the mean to replace missing values works effectively. To identify patients with cardiovascular disease, researchers used the UCI heart disease dataset in this research. To demonstrate their results, they have compared the effectiveness of various machine learning algorithms using the accuracy, precision, f1-score, and recall metrics. Using SVM with a linear kernel, a scoring accuracy of 86.8 percent overall was attained.

In [8] the researcher has managed to work on a wide range of machine learning such as the Random Forest algorithm, Support Vector Machine, Naïve Bayes, Logistic model tree algorithm, K-Nearest Neighbour, and data mining methods and has assessed them using the UCI heart disease dataset, which has 303 samples data with fourteen attribute values. Discovered that the SVM accuracy score of 84.1584 percent is the best among them; other algorithms include KNN, Naive Bayes, and decision tree.

In [9] This study Applies feature selection and classification algorithms, and a strategy for identifying cardiac disease has been presented. For feature engineering, feature selection techniques apply, and the Sequential Backward Selection Algorithm (SBS FS) is used. The Cleveland heart disease dataset was used to assess the model. The remaining 30% of the dataset was utilized for validating, and 70% for training. Utilizing assessment metrics, the suggested system's performances have been evaluated. On entire and selected feature sets, the classier K-Nearest Neighbor (K-NN) performance is evaluated. The suggested technique had a 90 percent prediction accuracy.

In [10], This paper's goal is to offer an optimization function based on a support vector machine (SVM). The genetic algorithm (GA) selects the most significant traits to develop heart disease using this objective function. the genetic algorithm is used to provide an efficient feature selection process. A dataset is taken from the Cleveland Heart Disease Database. The cardiovascular predictions are then created using a support vector classifier, which has an accuracy of 88.34 percent when diagnosing cardiac illness using the given attributes.

In [11], The purpose of this study is to find important characteristics and data mining methods that can increase the precision of heart disease prediction. Seven classification algorithms and various feature combinations were used to create prediction models such as Naive Bayes and Logistic Regression, k-NN, Decision Tree, Naive Bayes, Logistic Regression (LR), Support Vector Machine (SVM), Neural Network, and Vote approach. Datasets on cardiovascular disease were obtained from the UCI dataset repository. The researchers wanted to find a mix of variables and data mining approaches that might assist identify cardiovascular problems. To address the larger challenge of cardiac disease predictions, a comprehensive system framework that handles preprocessing, parameter tuning, and feature engineering is required. Results reveal that the diagnosis of the heart disease model created utilizing the recognized significant features and the top data mining method Vote achieves an accuracy of 87.4%.

In [12], To make a reliable prediction of heart disease, machine learning classifiers are created and a comparison analysis is done. Five ML algorithms are developed, and the Cleveland Heart Disease Data set is used to thoroughly assess each one's performance. These classifiers are Logistic Regression, Naive Bayes, Random Forest, Support Vector Machine, and K-Nearest Neighbor. After pre-processing, split the data set into training and testing halves using an 80/20 ratio. there are many well-known classification algorithms for detecting cardiac disease. To illustrate each model's efficiency in identifying cardiac disease, a comparison study of machine learning techniques is performed. Binary classifiers for pre-processed data have receiver operating characteristics that can be tuned using hyperparameters. LR achieved the highest accuracy, 0.93

In [13], To predict the presence of cardiac disease, this research suggests a three-layer, binary classifier based on neural networks (NNs). Univariate and bivariate exploratory data analysis were used in the filtering procedures that were used to create the feature space (EDA). In this study, the Cleveland UCI heart dataset was utilized. the artificially intelligent clinical decision systems at this time are currently gained so that medical efforts can be effectively saved to spend properly. The study makes use of several data engineering strategies to increase accuracy to a maximum of 91.66% and an average of 88.33%.

In [14], This research proposes a deep learning strategy for the diagnosis of heart illness based on Multiple Kernel Learning with an Adaptive Neuro-Fuzzy Inference System (MKL with ANFIS). the UCI heart disease characteristics are modeled using the Extreme Machine Learning Algorithm (EML). The proposed model achieves 80% of precision.

3. METHODOLOGY

The goal of this study is to create a model that can anticipate heart disease and optimize the classification result using one of the optimization methods. There is a portion in this part-is section that includes Data collection, dataset description, data pre-processing, feature engineering, and applicable machine learning algorithms, as well as a block diagram and evaluation matrices, and also the study's procedure and methodology. Fig.(1) shows the architecture of the proposed model.

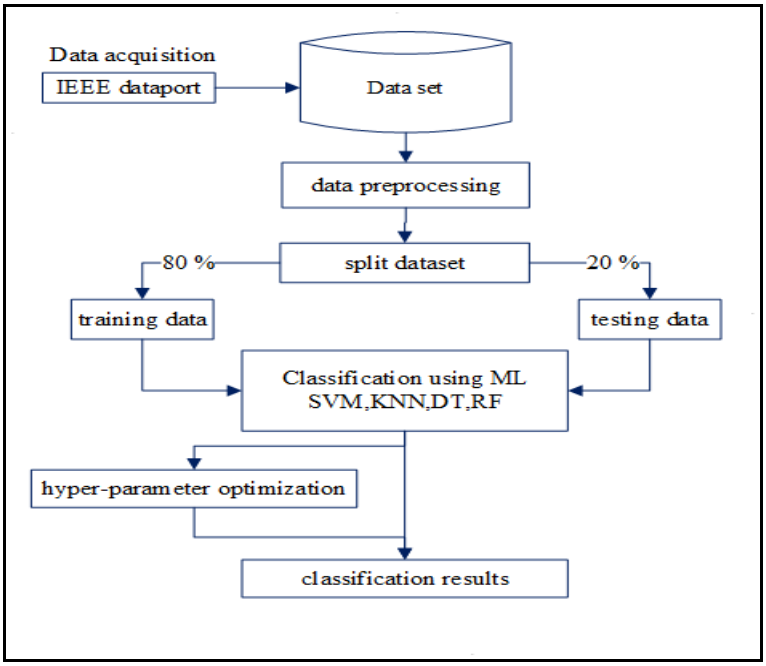

Figure (1): The architecture of the proposed model.

3.1 Dataset description

This cardiovascular disease dataset was compiled by integrating five famous cardiovascular disease datasets that had previously been available separately but had never been combined. This dataset includes five heart datasets with 11 related variables to give the most complete heart disease dataset for scientific study accessible. The five datasets that were utilized in its administration are as follows Long Shoreline VA, Hungarian, Cleveland, Starlog, Switzerland. There are 1190 samples in this dataset, each with 11 attributes. These datasets were gathered and pooled in one location to aid research into cardiovascular to enhance clinical diagnosis and early treatment in the future, therefore machine learning, and data mining approaches connected to illness CAD will be utilized. This data might be utilized to build a machine-learning algorithm model for detecting the earlier start of cardiovascular disease [15]. Table (1) shows the list of 11 traits on which the framework is working while Table (2) shows the description of a dataset of heart infections of minimal characteristics.

Table 1. - Attributes of the dataset

Attribute	Code given	Unit	Data type
Sex	Sex	1, 0	Binary
Age	Age	in years	Numeric
chest pain type	chest pain type	1,2,3,4	Nominal
resting blood pressure	resting bp s	in mm Hg	Numeric
exercise-induced angina	exercise angina	0,1	Binary
Serum cholesterol	cholesterol	in mg/dl	Numeric
old peak =ST	old peak	depression	Numeric
the slope of the peak exercise ST segment	ST slope	0,1,2	Nominal
resting electrocardiogram results	resting ECG	0,1,2	Nominal
fasting blood sugar	fasting blood sugar	1,0 > 120 mg/dl	Binary
maximum heart rate achieved	max heart rate	71–202	Numeric
Class	Target	0,1	Binary

Table 2. - Attributes of the dataset

Attribute	Description
Sex	1 = male, 0= female;
Class	1 = heart disease, 0 = Normal
Exercise-induced angina	1 = yes; 0 = no
the slope of the peak exercise ST segment	Value 1: upsloping, Value 2: flat, -- Value 3: downsloping
Resting Electrocardiogram results	Value 0: Normal, Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), Value 2: showing probable or definite left ventricular hypertrophy by Estes criteria
Fasting Blood sugar	(Fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

3.2 Data Pre-Processing

Data preparation is the most crucial initial stage in a somewhat analytical model; it assists to organize data in a readable fashion, which enhances model effectiveness. Medical data is frequently insufficient, missing characteristic values, and noisy owing to outliers or extraneous data [6].data preprocessing steps are :

1- Data cleaning: Data cleaning is a task in which information is cleaned by evacuating lost information, copying information, and settling information irregularities. As a result, data quality is progressing coming of the inconvenience of data [10]. We performed pre-processing on the information set, from a total of 1190 samples in this dataset 272 duplicate copy records were removed from the dataset. The remaining 918 patient records are used to identify whether or not a person has heart disease. The value is set to one of the patients who have heart disease, otherwise to zero, indicating that the patient does not have heart disease.

2- Outliers removal: The data in the dataset that differs significantly from the dataset's norm is an outlier. The bulk of the data's outliers are thought to be noise, which reduces the model's performance and doesn't add anything to the relevance of the data[16]. have shown that eliminating outliers from data aids in generating better outcomes and have developed several outlier removal strategies. Boxplot is used in our suggested framework to eliminate outliers. Boxplots are used to display the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and a maximum of five numerical values. Any point that lies outside of the box formed by these five points when they are plotted is regarded as an outlier[17]. In the cholesterol Colome also some outlier values are present so, in the same manner, will remove as shown in Fig.(2-a). An outlier the result of cholesterol after removing the outlier is shown in Fig.(2-b).

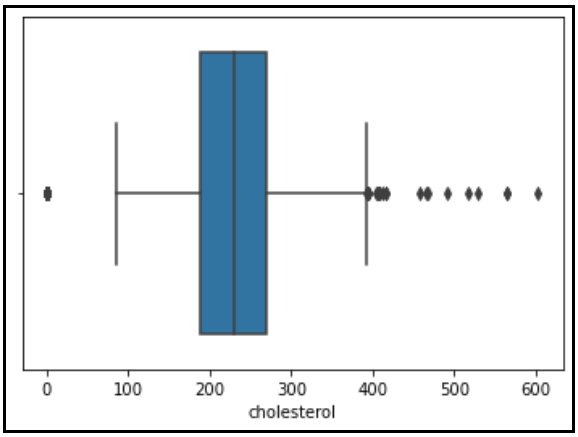

FIGURE 2. - **(a) cholesterol** before removing outlier

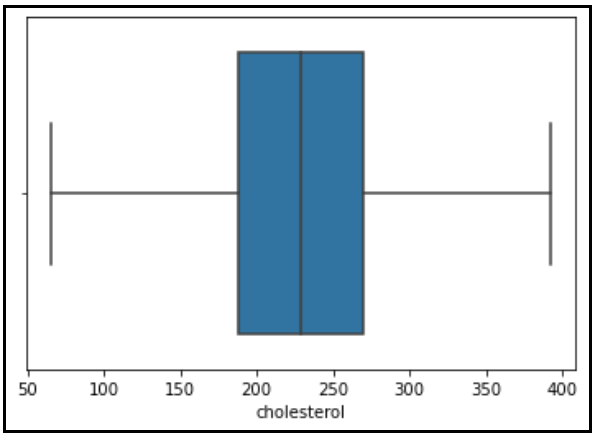

FIGURE 2. - **(b) cholesterol** after removing outlier

3- Data transformation: The modification of information or data from one format to another format is known as information change. It is ordinarily done when a source format is required to change over into the required format for a particular reason. It includes responsibilities for aggregation, standardization or normalization, and smoothing [18]. Normalization This Refers to rescaling a real numeric attribute to the range 0 to 1. Information normalization is utilized in machine learning to create a model prepared less delicate to the scale of highlights we utilize standard scaling methods[19].

3.4 Partitioning of data

Splitting is the process of randomly splitting the data set into two groups, the first of which will be used as training data and are going to be utilized as test data. A model is created and trained using training data, and then it is evaluated using testing data[20]. Based on this study, the researcher decided to divide the data into an 80:20 ratio, with 80 percent going to training and 20 percent going to validation.

3.5 Classification Analysis

The Algorithm that has been proposed for the most common disease in the healthcare field is heart disease, and its prevalence is rising yearly. The following three commonly employed machine learning algorithms for predicting heart disease were examined in this research, Random Forest, Support Vector Machines, KNN, and Decision Tree These methods are useful for identifying binary dependent variables. Ml learning algorithms used in this model are:

1- **Support Vector Machine:** supervised learning models which are used to analyze data and discover patterns in classification and regression analysis [21]. SVM is designed to discover hyperplanes in N-dimensional areas (N-features) that partition data. There are two ideas about whether or not data is linearly distinct given a dataset. The linear kernel works well if the data is linearly separable; however, if the data is not linearly distinct, it becomes difficult to separate the data. To perform classifying, a hyperplane is created, with samples from one class lying on one side and samples from another class lying on the other. To guarantee the greatest possible separation between the two classes, the hyperplane is optimized. Support vectors are those data points from classes that are closest to the hyperplane[22].

2- **K-Nearest Neighbor:** It is a categorization system based on distance measurements. It is an instance-based categorization, which implies that comparable instances are classified similarly. The slow or lazy algorithm is another name for it. We have the appropriate X-value and Y-value for each point. We receive the Y-value of both instances when we are given a new instance in terms of the X-value and find the Y-value of the instance. We want to be able to accurately estimate the majority class based on Yi values. We use distance functions like Euclidean Distance, Manhattan Distance, and Minkowski Distance to locate the most similar/ adjacent example[23].

3- **Decision Tree:** Decision trees (DTs), are one of the most powerful and widely used classifying and predictive methods in machine learning today. Many academics have utilized it as a classifier in the healthcare area to assess data and make choices. DT creates a model that predicts the value of a target variable by learning fundamental decision rules generated from data properties and splitting data into branch-like segments. There are two types of input values: continuous and discontinuous. The leaf nodes return class labels or probability scores. It is possible to turn the tree into a set of decision rules. These categorization principles can be readily shown graphically (6).

4- **Random Forest (RF):** is a data categorization strategy that employs a large number of decision trees. Bagging and feature randomization are used to create an uncorrelated forest of trees whose committee prediction is more accurate than any individual tree commonly used in classification and regression problems. To achieve the optimum outcome, this classification technique constructs many decision trees and combines them. It mostly uses bootstrap aggregation or bagging for tree learning[24].

3.6 Hyperparameter optimization

hyperparameters in the ML are used to control the operation of the algorithm in the model. Hyperparameter optimization fits a group of hyperparameters of the classification algorithm to enhance the operation of the ML model [25]. There are different types of hyperparameters in ML algorithms that must be tuned to improve the result [26].

3.6.1 Random Search

As candidate hyper-parameter values, RS chooses at random a predefined set of samples from the range between the upper and lower bounds. These candidates are subsequently trained up until the budget allotment is reached. According to the theory behind RS, the global optimums, or at the very least their approximations, can be found in the configuration space sufficiently big. Since each evaluation is independent, RS's key benefit is that it is simple to parallelize and distribute resources. It increases system efficiency by decreasing the likelihood of spending a lot of time on a small, underperforming region by sampling a fixed number of parameter combinations from the given distribution. Furthermore, if given enough budgets, RS can identify the global optimum or a close approximation of it [25]. The main steps in RS are :

Step 1: Best ← a few starting randomized potential solutions.
Step2 : Repeat
Step 3: S ← a potential randomized solution.
Step 4: if Accuracy (S) > Accuracy (Best) then
Step 5: Best ← S
Step 6: until Best is the optimal course of action or until time runs out.
Step 7: Return the Best

3.6.2 Hyperparameters used in Random Search optimization method

There are different types of Hyperparameters in each ML algorithm that must be tuned to enhance the result, Table (3) shows the Hyperparameters chosen for use in the RS optimization method with their definitions and the default values.

Table 3. - ML Hyperparameters used in Random Search

Classifier	Hyperparameter	Definition	Default
SVM	**C**	**Regularization parameter.**	**1**
	Gamma	**Kernel coefficient for 'rbf', 'poly', and 'sigmoid'.**	**Scale**
KNN	**n-neighbors**	**Number of neighbors**	**5**
	Weights	**Weight function used in prediction.**	**uniform**
	Metric	**Metric to use for distance computation**	**Minkowski**
DT	**Max- depth**	**The maximum depth of the tree.**	**None**
	Criterion	**The function to measure the quality of a split**	**Gini**
	Max- features	**he number of features to consider when looking for the best split**	**None**
	Min-samples-split	**The minimum number of samples required to be at a leaf node**	**1**
RF	**n_estimators**	**The number of trees in the forest**	**100**
	max_features	**he number of features to consider when looking for the best split**	**sqrt**
	max_depth	**The maximum depth of the tree.**	**None**
	Min-samples-split	**The minimum number of samples required to split an internal node**	**2**
	min_samples_leaf	**The minimum number of samples required to be at a leaf node**	**1**
	bootstrap	**Whether bootstrap samples are used when building trees**	**True**

3.7 Performance measurement

the performance of many classification methods must be evaluated, the evaluation metrics are important to estimate both the performance and result of the classification. The evaluation technique conducted utilized a confusion matrix also called a contingency table.

4) shows the equations that are used for evaluating the classifiers' performance [27].

Table 4. - Evaluation metrics

Metric Name	Calculation
Recall	$TP/(TP+FN)$
Precision	$TP/(TP+FP)$
F1 Score	$2 * (Precision * Recall)/(Precision + Recall)$
Accuracy	$(TP+TN)/(TP+FN+TN+FP)$

4. Results and Discussion

Before data preprocessing, the dataset must split into (80- 20) % train-test dataset. This splitting ratio gives the ability to model for train and testing on an unseen dataset and the train and test set is then applied to the classifiers in the mode. Table (5) shows the classification result using (SVM, KNN, DT, and RF) classifiers and the best results when using the RF classifier, while Fig.(3) shows the comparison of test accuracy results for all classifiers without and with Standard Scaling

Table 5. - **Classification results**

Classifiers	Acc without Scaling	Acc + Standard Scaling	Precision	Recall	Specificity	F1- score
SVM	72.68	89.0	0.87	0.93	0.83	0.90
KNN	71.84	88.6	0.87	0.92	0.84	0.89
DT	89.91	89.9	0.93	0.86	0.92	0.87
RF	***94.53***	***94.9***	***0.94***	***0.96***	***0.93***	***0.95***

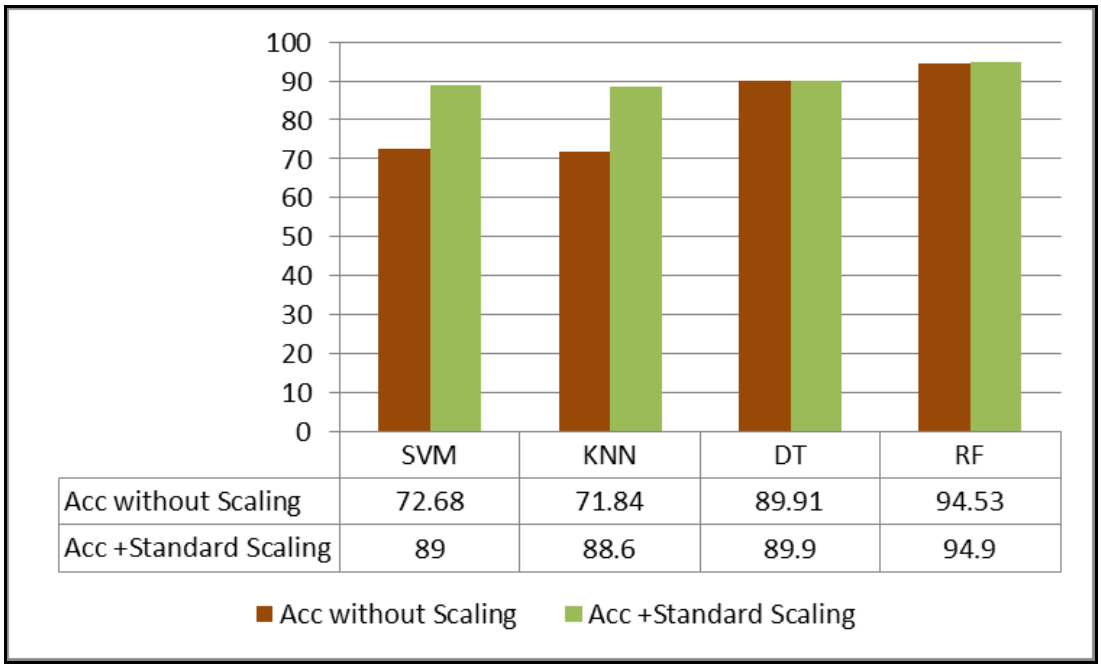

Figure (3): Best Acc comparison Without and with Standard Scaling

After completing the classification stage and before the optimization stage, hyperparameter ranges for all classifiers must be selected as shown in Table (6).

Table 6. - **Classifiers Hyperparameter values in experiments**

Classifier	Hyperparameter	Values used in experiments
SVM	C	-10 to 21
	Gamma	['scale', 'auto']
KNN	n-neighbors	(5, 7, 11, 13, 15)
	Weights	(Uniform, distance)
	Metric	(Minkowski, Euclidean, manhattan)
DT	Max- depth	(200,150,100,50,25)
	Criterion	(Gini, entropy)
	Max- features	(auto, sqrt, log2)
	Min-samples-split	(2,4,6)
RF	n_estimators	(start = 1, stop = 20, num = 20)
	max_features	['auto', 'sqrt', 'log2']
	max_depth	(100, 10000, num = 12)
	Min-samples-split	[2, 6, 10,14,16,20]
	min_samples_leaf	[1, 3, 4,5,6]
	Bootstrap	[True, False]

The optimization stage begins with applying RS to the data set with all classifiers, and this method was used to select the optimal value for the parameters specified for each classifier that affects the classification result. Table (7) shows the best test accuracy with the optimal value of hyperparameters for each classifier and the best result when using the RF classifier while Fig. (4) shows the best Acc comparison before and after optimization and table (8) shows Compression between this work and previous works.

Table 7. - **Optimization result**

Classifier	Hyperparameter	Optimal values	Best Acc
SVM	C	10	89.0
	Gamma	Auto	
KNN	n-neighbors	15	93.0
	Weights	distance	
	Metric	manhattan	
DT	Max- depth	15	94.0
	Criterion	entropy	
	Max- features	auto	
	Min-samples-split	4	
RF	n_estimators	9	95.4
	max_features	Log2	
	max_depth	1000	
	Min-samples-split	16	
	min_samples_leaf	3	
	bootstrap	False	

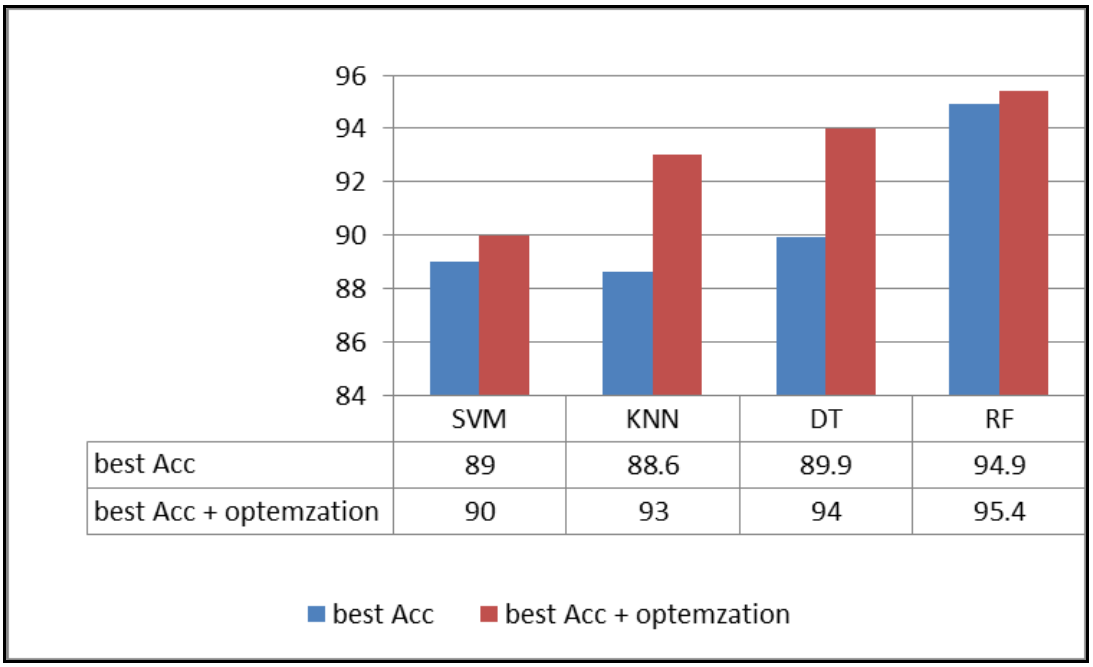

Figure (4): Best Acc comparison before and after optimization

Table 8. - Compression of this work with previous works

Studies	Classifier	best accuracy
[7]	SVM	86.8
[8]	SVM	84.15
[9]	KNN	90.0
[10]	SVM	88.34
This work		
without optimization	SVM	89.0
	KNN	88.6
	DT	89.9
	RF	94.9
with optimization	SVM	90.0
	KNN	93.0
	DT	94.0
	RF	95.4

5. Conclusions

The heart is one of the most vital organs in the human system. This paper presents a proposed model to design and implement an automated model to predict heart attacks with high accuracy in the early stages. Data preprocessing techniques and machine learning algorithms were used to achieve the highest desired efficiency of the model. The model validation is conducted with the train-test split (80-20) of the dataset. The experiment result revealed that RF achieved better accuracy than all comparative models as shown in Table (5).

After the start of the optimization process, the parameters for each classifier and its range are determined using the RS optimization method, and the highest results were obtained when using the RF classifier with max depth(1000), max_features (Log2), n_estimators (9), Min-samples-split(16), min_samples_leaf (3) and bootstrap (False) .all these optimal values enhance the result and reached the best test Acc to (95.4 percent)

Funding

None

ACKNOLEDGEMENT

The author we would like to thank the reviewers for their valuable contribution in the publication of this paper.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] K. D. 1K. Venkatesh1, M. Prathyusha3, C.H. Naveen Teja4, "Identification of Disease Prediction Based on Symptoms Using Machine Learning," *JAC : A Journal Of Composition Theory*, vol. XIV, no. VI, June 2021. [Online]. Available: http://www.jctjournal.com/gallery/10-june2021.pdf.

[2] M. S. Ricardo Buettner, "Efficient machine learning based detection of heart disease," presented at the 2019 IEEE International Conference on E-health Networking, Application & Services (HealthCom), 2019, 1. [Online]. Available: https://ieeexplore.ieee.org/document/9009429.

[3] R. K. Archana Singh, "Heart Disease Prediction Using Machine Learning Algorithms," presented at the 2020 International Conference on Electrical and Electronics Engineering (ICE3), 26 June 2020, 2020, 1. [Online]. Available: https://ieeexplore.ieee.org/document/9122958.

[4] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," *IEEE Access*, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/access.2020.3001149.

[5] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Computing and Applications*, vol. 29, no. 10, pp. 685-693, 2016, doi: 10.1007/s00521-016-2604-1.

[6] S. B. Halima El Hamdaoui, Nour El Houda Chaoui, Mustapha Maaroufi, "A Clinical support system for Prediction of Heart Disease using Machine Learning Techniques," presented at the 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), 02-September 2020, 2020, 1. [Online]. Available: https://ieeexplore.ieee.org/document/9231760.

[7] M. A. Nabaouia Louridi, Bouabid El Ouahidi, "Identification of Cardiovascular Diseases Using Machine Learning," presented at the 2019 7th Mediterranean Congress of Telecommunications (CMT), 2019, 1. [Online]. Available: https://ieeexplore.ieee.org/document/8931411.

[8] K. V. s. K. Nikhil Kumar Muthyala, K.Deepak, "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools," *IJSRCSEIT*, vol. 3, no. 3, p. 13, 15 March 2018 2018, doi: 10.13140/RG.2.2.28488.83203.

[9] J. L. Amin ul Haq, Muhammad Hammad Memon, Muhammad Hunain Memon, Jalaluddin Khan, Syeda Munazza Marium, "Heart Disease Prediction System Using Model Of Machine Learning and Sequential Backward Selection Algorithm for Features Selection," presented at the 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 29 -31 March 2019, 2019, 1. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9033683.

[10] C. B. Gokulnath and S. P. Shantharajah, "An optimized feature selection based on genetic approach and support vector machine for heart disease," *Cluster Computing*, vol. 22, no. S6, pp. 14777-14787, 2018, doi: 10.1007/s10586-018-2416-4.

[11] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telematics and Informatics*, vol. 36, pp. 82-93, 2019, doi: 10.1016/j.tele.2018.11.007.

[12] H. S. Hiren Kumar Thakkar, Sonali Patil, "A Comparative Analysis of Machine Learning Classifiers for Robust Heart Disease Prediction," presented at the 2020 IEEE 17th India Council International Conference (INDICON, 2020, 1. [Online]. Available: https://ieeexplore.ieee.org/document/9342444.

[13] T. H. Shiti Maitra, Abdullah Al-Sakin, Faisal Muhammad Shah, "Artificial Prognosis of Cardiac Disease using an NN A Data-scientific Approach in Outlier Handling," presented at the 2019 4th International Conference on Electrical Information and Communication Technology (EICT), 20 Dec 2019, 2019, 1. [Online]. Available: https://ieeexplore.ieee.org/document/9068847.

[14] G. Manogaran, R. Varatharajan, and M. K. Priyan, "Hybrid Recommendation System for Heart Disease Diagnosis based on Multiple Kernel Learning with Adaptive Neuro-Fuzzy Inference System," *Multimedia Tools and Applications*, vol. 77, no. 4, pp. 4379-4399, 2017, doi: 10.1007/s11042-017-5515-y.

[15] N. Manu Siddhartha, 2020, "Heart Disease Dataset (Comprehensive)", IEEE Dataport, doi: https://dx.doi.org/10.21227/dz4t-cm36. [Online]. Available: https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive

[16] M. D. A. Pranatha, N. Pramaita, M. Sudarma, and I. M. O. Widyantara, "Filtering outlier data using box whisker plot method for fuzzy time series rainfall forecasting," in *2018 4th International Conference on Wireless and Telematics (ICWT)*, 2018: IEEE, pp. 1-4.

[17] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, "An integrated machine learning framework for effective prediction of cardiovascular diseases," *IEEE Access*, vol. 9, pp. 106575-106588, 2021.

[18] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Computer Science*, vol. 1, no. 6, 2020, doi: 10.1007/s42979-020-00365-y.

[19] A. A.-M. Rahma Ataullah, "Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method," vol. 1, p. 6, 2019.

[20] A. A. Ali, "Stroke Prediction using Distributed Machine Learning Based on Apache Spark," *Stroke*, vol. 28, no. 15, pp. 89-97, 2019. [Online]. Available: https://www.researchgate.net/profile/Nahla-Omran-2/publication/338458550_Stroke_Prediction_using_Distributed_Machine_Learning_Based_on_Apache_Spark/links/5e1619404585159aa4be6a2e/Stroke-Prediction-using-Distributed-Machine-Learning-Based-on-Apache-Spark.pdf.

[21] V. Sharma, S. Yadav, and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," presented at the 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9137817.

[22] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," *Journal of Reliable Intelligent Environments*, vol. 7, no. 3, pp. 263-275, 2021, doi: 10.1007/s40860-021-00133-6.

[23] P. A. T. Azhar M.A, "Comparative Review of Feature Selection and Classification modeling," presented at the 2019 International Conference on Advances in Computing, Communication and Control (ICAC3), 7 Mar 2019, 2019, 1. [Online]. Available: https://ieeexplore.ieee.org/document/9036816.

[24] Sabah Abdul kareem, H., & Mahdi Altaei, M. S. (2023). Detection of Deep Fake in Face Images Based Machine Learning. Al-Salam Journal for Engineering and Technology, 2(2), 1–12. https://doi.org/10.55145/ajest.2023.02.02.001

[25] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295-316, 2020.
[26] M. Feurer and F. Hutter, "Hyperparameter optimization," in *Automated machine learning*: Springer, Cham, 2019, pp. 3-33.
[27] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text classification using machine learning techniques," *WSEAS transactions on computers*, vol. 4, no. 8, pp. 966-974, 2005.