Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Adaptive Features Selection Technique for Efficient Heart Disease Prediction

Zahraa Ch. Oleiwi [a]*, Ebtesam N. AlShemmary [b], Salam Al-augby[c]

[a]*College of Computer Science and Information Technology, University Al-Qadisiyah.Email: zahraa.chaffat@qu.edu.iq*

[b]*IT Research and Development Center, University of Kufa, Najaf, Iraq.Email: dr.alshemmary@uokufa.edu.iq*

[c]*Faculty of Computer Science and Mathematics, University of Kufa, Najaf, Iraq.Email: salam.alaugby@uokufa.edu.iq*

ARTICLE INFO

Article history:
Received: 15 /11/2022
Revised form: 03 /01/2023
Accepted: 09 /01/2023
Available online: 17 /02/2023

Keywords:

Heart disease;

Features selection;

Mutual information;

Random Forest

ABSTRACT

Heart disease is a common disease that causes death and is difficult to detect manually. A more efficient classification model that relies on machine learning methods to achieve higher classification accuracy, attracts the attention of researchers to design an effective prediction model. Moreover, it plays an important role in the practical application of medical cardiology with the aim of early detection of heart diseases. In this paper, an efficient and accurate heart disease detection system is proposed based on the proposed adaptive feature selection technique using four machine learning methods: Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). Two feature selection methods were used to design the proposed technique, mutual information (MI) and recursive feature elimination (RFE) to determine the optimal number of selected features that increase the performance of the classification models and reduce the time complexity of model implementation. The proposed technique was implemented on the two standard databases from the UCI machine learning repository: Cleveland heart disease and heart Statlog Cleveland. The best model was selected and saved as a prediction model using the cross-validation method. The results show that each data has a different number of features chosen according to the classifier model. For the first heart disease dataset, the best heart disease detection system Support Vector Machine-mutual information (SVM-MI) achieved the highest classification accuracy of approximately 96.755 compared to the other classifier models used. While the Random Forest-mutual information (RF-MI) model achieved an accuracy of 97.4% for the second data set. The proposed technique produced the highest prediction performance in terms of accuracy, f1 score, accuracy, and metric retrieval compared to the latest research in this field.

MSC.

https://doi.org/10.29304/jqcm.2023.15.1.1137

∗Corresponding author

Email addresses: zahraa.chaffat@qu.edu.iq

Communicated by 'sub etitor'

Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

1. Introduction

According to the European Society of Cardiology, approximately 26 million people are suffering from heart disease, which is considered a serious and life-threatening disease[1]. As a result, the design of effective and efficient techniques for the accurate and early detection of these diseases has become one of the most important topics that most researchers are looking for, as this has an impact on saving the lives of many patients[2], where the manual diagnosis has many difficulties and problem in terms of time-consuming and inaccurate diagnosis.

Consequently, to avoid the difficulty of manual and inaccurate heart disease diagnosis, many automatic classification techniques based on machine learning methods were designed such as Logistic Regressing (LR)[3], Naive Bayes (NB)[4], Support Vector Machines (SVM) [5], Neuro-Fuzzy approach[26], Linear Discriminant Analysis (LDA), The K-Nearest Neighbor (K-NN)[5], Decision Tree (DT)[5], neural networks[5], and Random Forest (RF)[6].

Moreover, the trend of all research aims to increase the performance of automatic classification techniques to be suitable for the sensitivity of medical applications for time and accuracy. Many factors affect the performance of classification model-based machine learning, one of these factors is the characteristics of the dataset: accuracy and balancing dataset, size, and dimensionality of the dataset, sparsity, importance, and relevance of features. So, feature selection is an important and essential step required to improve the performance of the classification model in terms of decreasing the complexity (fast convergence) and increasing accuracy (avoiding overfitting)[7, 8].

There are three types of feature selection methods classified according to the way of selecting the features relevant that maximize the classification decision accuracy: filter(rank), wrapper, and embedded-based selection method. For each type of feature selection there are advantages and disadvantages, therefore, choosing the suitable selection method depends on the dataset's characteristics and the application of the model implemented [9].

Filter-based feature selection methods select features by ranking features according to their importance, relevance, and correlation with the target class independent of the classifier algorithm. The ranking methods used to evaluate the feature are: information (using information theory concepts) as in mutual information feature selection method, statistical such as correlation and Chi-square based selection method, similarity measure as in Spectral feature selection (SPEC) and Laplacian Score (LS), and distance measure as in Relief method[10].

Wrapper-based feature selection methods select a subset of features by employing the classifier to choose the features having higher estimation methods i.e., features score the highest classification accuracy. Support Vector Machine Recursive Feature Elimination (SVM-RFE) is the most common wrapper feature selection method. Filter-based feature selection methods are characterized by less complexity and efficiency than wrapper-based methods, while wrapper-based methods achieve higher accuracy than filter-based methods[10].

In literature, different commonly used techniques of feature selection are such as Relief, Principal component Analysis (PCA), Greedy Algorithm (GA), Minimal-Redundancy-Maximal-Relevance (MRMR), Least-absolute-shrinkage-selection-operator (LASSO). Other feature selection methods are optimization methods such as fruit fly optimization

∗Corresponding author

Email addresses: zahraa.chaffat@qu.edu.iq

Communicated by 'sub etitor'

(FFO), Ant Conley Optimization (ACO), Particle Swarm Optimization (PSO), Bacterial Foraging Optimization (BFO), and other methods, which are among the hot topics of modern research[1, 2].

All features selection methods require the common input parameter that refers to the required number of selection features (size of subset feature selected vector) and this parameter change according to the type of dataset and classifier. On the other hand, it is not reasonable to try all features, especially for datasets with large dimensions. In addition, some datasets do not need to feature selection since there is underfitting in classifier performance[10].

In this context, the work in this paper aims to produce an adaptive best number of selected features for any dataset, any feature selection method, and classifier with aim of finding the best classifier model with less than the most useful number of features for two heart disease datasets. On the other hand, the general aim of this work is to develop an efficient heart disease classification framework in terms of less complexity and high accuracy.

To achieve and investigate these aims the research work has the following objectives illustrated in bellow steps:

1- Two heart diseases are preprocessed to implement the proposed work.
2- Building a function that returns the best number of features by utilizing two feature selection methods: RFE and Mutual information, and four machine learning classifiers: logistic regression, K-nearest neighbor, Decision tree, and Random Forest.
3- Constructing a new dataset including k best features. Creating a function for model cross-validation evaluation by employing a k-fold algorithm. This function takes a new dataset and different classifier model as input and returns model evaluation metrics for each model with indices of train set and test set at k the model achieves high performance with it
4- Select and save the best classifier model for each dataset to use for heart disease prediction to predict whether a new test instance has heart disease or not.

The rest of this paper is organized as follows. First, the related work is described in Section 2. Section 3 describes the methodology of the work. Section 4 discusses the results and analysis of the proposed work. Finally, the conclusions and future work are presented in Section 5.

2. Related works

Recently, most of the techniques based on machine learning methods are proposed to design heart disease prediction systems and all these techniques aim to increase the accuracy and efficiency of classification since the medical application is sensitive to time and accuracy factors. Most of these techniques depend on developing feature selection methods as trials to achieve more efficiency and accuracy.

A heart disease diagnosis system was proposed in [1]by using six classification methods: K-nearest neighbor (KNN), Artificial neural network (ANN), Decision tree (DT), Support vector machine (SVM), Logistic regression (LR), and Naïve bays (NB), and four features selection methods: Minimal redundancy maximal relevance (MRMR), Relief, Local learning, and Least absolute shrinkage selection operator (LASSO) in the aim of reducing the time complexity and achieve efficient and more accurate classification system. In this work, the author proposed a new feature selection method called the fast conditional mutual information feature selection algorithm (FCMIM). A comparative and analytical study was produced for all classification and feature selection methods. As a result, the FCMIM with the support vector machine (SVM) achieves higher accuracy about 92% with seven selected features from the original Cleveland Heart Disease dataset.

A comparative study on the analysis of different machine learning-based heart disease classification techniques was presented in [11]where different machine learning methods DT, LR, SVM, NB, ANN, and a hybrid model with LR and BN were used in this study. the results in this study show that the hybrid model with selected features achieves higher classification accuracy about 87.41% as compared with another used classifier.

Seven machine learning methods-based heart disease classification: Support Vector Machine (SVM), k-NN, Naive Bayes, Decision Tree, Logistic Regression (LR), Neural and Vote networks were employed IN [8] to identify the best

subset features and best classifier of cardiac diseases predict. the outcomes of this research achieve a precision of about 91.4% and the authors concluded that the high impact of using relevant features with a strong analysis of data on obtaining high performance in terms of classification precision.

Another study proving the impact of preprocessing, analysis, understanding, and improving the quality of the heart diseases dataset on improving the performance of classification techniques was produced in [9]. The performance of different classification methods: K-nearest neighbor, logistic regression, Gaussian naive Bayes, decision tree, random forest, and support vector machine was examined on the Cleveland heart diseases dataset with features selected using ten features selection methods: recursive feature elimination (RFE), forward feature selection, ReliefF, Lasso regression, Ridge regression, ANOVA, Chi-square, mutual information, backward feature selection, and exhaustive feature selection. as a result, the outcomes of this research show that the highest accuracy of 88.52% was obtained by using the decision tree classifier which was trained on a subset of features selected by the backward feature selection method.

Although all the above-related research produced beneficial feature selection techniques, they have some drawbacks. some of them do not achieve the required accuracy. No of these researches identified the optimal number of features without using trying ways. in this paper adaptive technique is used to identify the optimal relevant features that are suitable to any datasets and classifiers in an adaptive way that produces efficient and more accurate classification performance.

3. Material and Methods

All the theoretical backgrounds of methods and materials used in this research are explained in this section.

3.1 Datasets Description

The brief explanation and website of the two used heart disease datasets available in it are produced as follows:

- The first Cleveland heart disease database contains 303 instances and 76 attributes but only 14 of them are used as standard in most studies, including the predicted attribute. It is available at (https://archive.ics.uci.edu/ml/datasets/heart+disease)[12].
- The second heart Statlog Cleveland Hungary database is gathered from the UCI machine learning repository and combined from five different datasets: Cleveland: 303, Hungarian: 294, Switzerland: 123, Stalog (Heart) Data Set: 270, and VA Long Beach: 200, thus, it consists of 1190 instances (patients from US, UK, Switzerland, and Hungary) and 11 common attributes[2]. It is available at (https://www.kaggle.com/datasets/sid321axn/heart-statlog-cleveland-hungary-final).

3.2 Datasets preprocessing

Data preprocessing is a significant step to acquire good quality data that impact the performance of the general model. The two used datasets required scaling using Standard Normal Distribution (SND). As a result, the resultant transformed scaled data has been distributed with zero mean and unit variance.

3.3 Mutual Information-based Feature Selection Method (MI-FS)

The mutual information (MI) measurement is one of the information theory concepts[13]. For two random variables X and Y, MI used as measurement that measure the relevance between them by measuring the information amount of X contains in Y and the information amount of Y contains in X.

Definition 3.1 (Entropy) Entropy is a measure of the average level of uncertainty or information of outcome possible for a random variable. For random variable X if $p(x)$ is the probability of X then the entropy denoted by $H(X)$ as in equation (1):

$$H(X) = -\sum_{x \in X} p(x) \log\big(p(x)\big) \qquad (1)$$

For the two variables X and Y with joint probability $p(x, y)$, the joint entropy is denoted by $H(X, Y)$ as equation (2)

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log\big(p(x,y)\big) \qquad (2)$$

On other hand, the entropy of a variable given (condition by) another variable is denoted by $H(X|Y)$ as in equation (3):

$$H(Y|X) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log\big(p(y|x)\big) \qquad (3)$$

Definition 3.2 (Mutual Information) if the joint probability between the variables X and Y is $p(x,y)$, $p(x)$ is the marginal probability density functions for X, and $p(y)$ is the marginal probability density functions for Y, then the MI for X and Y is $I(X;Y)$ as equation (4):

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)log\frac{p(x,y)}{p(x)p(y)} \qquad (4)$$

i.e., MI is the relative entropy between $p(x,y)$ and the product $p(x)p(y)$

Definition 3.3 (Conditional Mutual Information) is the MI between the two random variables X and Y given Z as equation (5):

$$I(X;Y|Z) = \sum_{z\in Z}\sum_{y\in Y}\sum_{x\in X} p(x,y,z)log\frac{p(z)p(x,y,z)}{p(x,z)p(y,z)} = H(X|Z) - H(X|Y,Z) \qquad (5)$$

So, mutual information can be written as a relation between it and entropy as Equation (6):

$$I(X;Y) = H(X) - H(X|Y) \qquad (6)$$

Relevancy and redundancy are two terms the filter-based feature selection methods depend on. Where the feature selection method aims to find and select features that are more relevant to the target class and remove redundant features (the features that depend on or correlate with each other).
According to equations (5) and (6) the mutual information can measure the relevancy between the features X_m and class C as:
If $I(C;X_m) = 0$ then according to the equation (6) $H(X) = H(X|Y)$ so, X_m is independent to target class C and has no relevant information on classification

If $I(C;X_m) > 0$ then X_m is the relevant feature and its classification information increase with increasing of MI.

In addition, redundancy can be measured using MI by measuring the MI between features where the features that have high MI consider redundant variables.

In the light of above, mutual information-based feature selection methods select features with highly important classification i.e., features with high MI with target class, and remove redundant and irrelevant features. According to this mechanism in selecting features mutual information-based features selection methods is one of the filter-based features selection methods [13].

3.4 Recursive Feature Elimination-based Feature Selection Method (RFE-FS)

This method is based on the mechanism of eliminating features with less importance repeatedly until obtained the required number of selected features is.
At each iteration, the model will be trained and fewer important features removed. the significance of features is calculated depending on the weights of the algorithm for each iteration [10].
Initially, two parameters are required, the number of selected features and the estimator to be trained. the estimator is trained to all features in the original data to determine the significance of each feature then the less significant features are eliminated until obtain to the required number of features which is set in the initial step [9].

3.5 Classifier-based machine learning methods

Four machine learning are used in this work: logistic regression, Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM), the detailed explanation of these methods is in the following references [3, 5, 6]. The parameters of each classifier used in this work are clarified in Table (1).

Table 1 – parameters setting for each machine learning model.

Machine learning model	Parameters
Logistic regression	Penalty=l2; C: Inverse of regularization strength=1; Algorithm to use in the optimization problem= 'lbfgs'
Decision Tree	The function to measure the quality of a split=Gini; The strategy used to choose the split at each node= best;
Random forest	The number of trees=50; The function to measure the quality of a split=Gini
Support vector machine	Regularization parameter=20; kernel=rbf; gamma (Kernel coefficient) =1

The rest of the parameters are set as in the default setting in Python's built-in functions

3.6 General Framework of Proposed Heart Disease Prediction Methodology

Adaptive best feature selection and general step of this work will be explained in this section as in Figure 1.

3.6.1 Adaptive Getting the best number of features

To avoid entering manually the number of features to be selected using features selection methods, the adaptive and automatic method of choosing the optimal number of features that maximize the evaluation metrics of classification mode performance will be produced as in Algorithm (1). This function tries to input all possible number of features as input to the feature selection method and then trains the model with each extracted new subset feature by employing the k-fold method for dataset splitting. Finally, model evaluation metrics are computed for each new subset and then the number of selected features that achieve the highest evaluation metrics in terms of accuracy, recall, precision, and fi-score, will be returned. As a result, the new subset will be constructed from the original training and testing dataset with the optimal number of features.

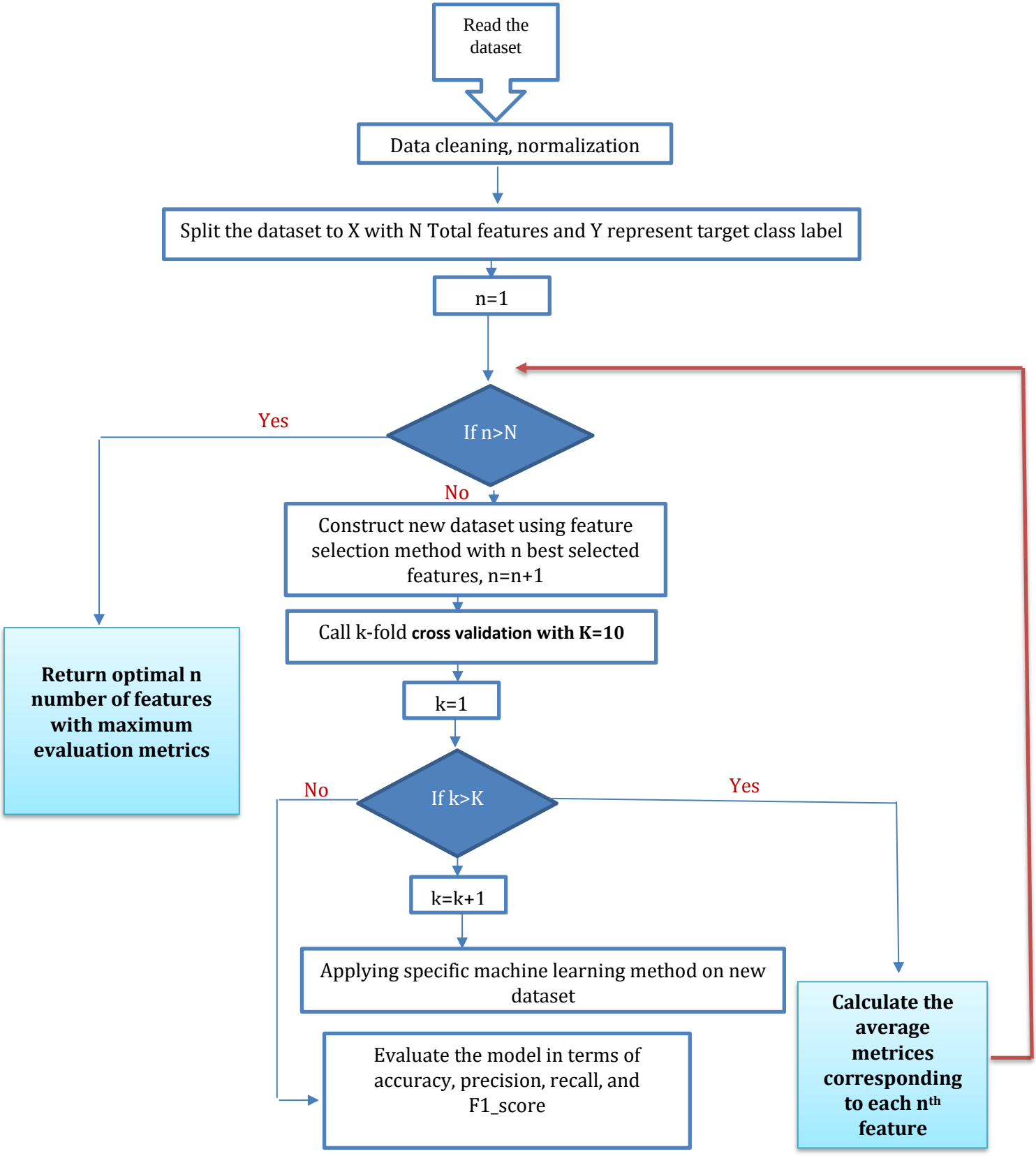

Fig. 1 – The general proposed framework for the adaptive feature.

ALGORITHM 1: GET THE BEST K FEATURES

Input: The heart Disease dataset features columns as x, target class label column as y

Output: best number of selected features as k

```
# Determine mutual information between features and target class
model evaluation ← {'f1': [], 'acc': [], 'k': []}
cols ← number of features in the dataset
for i:1to cols do:
    f1_score← [], acc ← []
    model ← classifier # e.g. ensemble.RandomForestClassifier
    sel_i_featuers ← feature selector(k=i)
    sel_i_ featuers.fit(x, y)
    new_x ← sel_i_ featuers.transform(x)
    kfold ← call(model_selection.StratifiedShuffleSplit)
    for train_indices,test_indices in kfold.split(new_x,y)do:
        x_train, x_test,y_train,y_test←kfold_train_test_split(new_x,y,train_indices,test_indices)
        model.fit(x_train,y_train)
        pred←model.predict(x_test)
        f1_score.append(metrics.f1_score(y_test,pred))
        acc.append(metrics.accuracy_score(y_test,pred))
    end
    model_evaluation['acc'].append( np.mean(acc))
    model_evaluation['f1'].append(np.mean(f1_score))
    model_evaluation['k'].append(i)
end
info←pd.DataFrame(model_evaluation)
k← info[info.f1 == info.f1.max()].k.values[0]
Return info, k # return the model evaluation data frame
```

3.6.2 The Model Cross-Validation Evaluation

This function enhances the performance of the model by controlling and avoiding the overfitting problem. By utilizing the k-fold method with k=10, the function is designed as in Algorithm (2). This algorithm takes the newly constructed dataset and model as input and returns the average of evaluation metrics (accuracy, recall, precision, and f1-score), evaluation metrics at each k, training set, and the testing dataset at k with maximum evaluation metrics.

ALGORITHM 2: MODEL CROSS VALIDATION EVALUATION

Input: The heart Disease dataset features columns as x, a target class label column as y, a classifier model

Output: average of evaluation metrics, evaluation metrics at each k, training set, and the testing dataset at k with maximum evaluation metrics.

```
1  model_evaluation ←{'f1':[], 'acc':[], 'precision':[], 'recall':[]}
2  kfold ← call(model_selection.StratifiedShuffleSplit)
3  max←0
4  for train_indices,test_indices in kfold.split(new_x,y)do:
5      x_train, x_test, y_train, y_test←kfold_train_test_split(new_x, y,train_indices,test_indices)
6      model.fit(x_train,y_train)
7      pred←model.predict(x_test)
8      if max<metrics.accuracy_score(y_test,pred):
9          max←metrics.accuracy_score(y_test,pred)
10         xtrain←x_train, ytrain←y_train
11         xtest←x_test, ytest←y_test
       end
12     model_evaluation['acc'].append(metrics.accuracy_score(y_test,pred))
13     model_evaluation['f1'].append(metrics.f1_score(y_test,pred,))
14     model_evaluation['precision'].append(metrics.precision_score(y_test,pred))
15     model_evaluation['recall'].append(metrics.recall_score(y_test,pred))
   end
16 model_evaluation ←pd.DataFrame(model_evaluation)
17 return model_evaluation, xtrain, ytrain ,xtest, ytest
```

3.6.3 Proposed Heart Disease Prediction Technique

The procedure we will work with to find the best model can be summarized in bellow pseudocode steps:

Step1: ***Read the data***

df←read_csv(dataset file)

Step2: ***Split pdf to features and target***

x ←df[:,:-1]

y← df[:,-1]

Step3: ***Preprocess dataset***

scaler ← preprocessing.StandardScaler()

scaler.fit(x)

scaled_x ← scaler.transform(x)

Step4: ***Define k-fold function***

```
def  kfold_train_test_split(x,y,train_indices,test_indices):
   return x[train_indices],x[test_indices] ,y[train_indices], y[test_indices]
```

Step5: ***Call Adaptive function of getting best k features (Algorithm (1))***
K←optimal number of features

Step6: ***Construct a new dataset of the optimal number of features***
new_x←optimal subset of the feature vector.transform(x)

Step7: ***For each model:***
Call Model Cross Validation Evaluation Function (Algorithm (2))
new_xtrain← x_train, new_ytrain← y_train
new_xtest← x_test, new_ytest← y_test

Step8: ***Test each model with new dataset***
prediction = model. Predict(new_xtest)
accuracy←metrics. Accuracy_score (new_ytest, prediction)

Step9: *B model←best classification model*

Step 10: *Save the best model*

Step11: ***Train and test the data with best model***
best_model .fit(new_xtrain, new_ytrain)
prediction = best_model. Predict(new_xtest)
accuracy←metrics. Accuracy_score (new_ytest, prediction)

Finally, the saved best model can be used as a prediction algorithm to predict the new patient sample whether has heart disease or not.

4. Results and Discussion

4.1 Experimental results

The proposed classification technique with adaptive feature selection algorithms using MI and RFE methods has been implemented on two standard heart disease databases. The results of this implementation using four classification methods are illustrated below in Tables.

The results in a Table (2) and (4) are obtained by calculating the maximum evaluation metrics after employing the k-fold cross-validation on the first and second datasets respectively with k=10. While the results in Tables (3) and (5) are obtained by calculating the average of evaluation metrics.

Tables (2) and (3) show the number of selected features from the original first Cleveland heart disease database containing 303 instances and 13 attributes, as well as the performance of each classifier used in terms of accuracy, f1-score, precision, and recall.

Table 2 – Maximum evaluation measures at specific k-fold cross-validation corresponding to the number of selected features and different classifiers implemented on the first dataset.

Classifier model	MI						RFE					
	No. of features	k-fold cross validation	Accuracy	F1-score	precision	recall	No. of features	k-fold cross validation	Accuracy	F1-score	precision	recall
LR	7	9	90.3%	90.9%	93.7%	88.2%	8	3	93.5%	94.4%	89.4%	100%
DT	**3**	**9**	**93.5%**	94.1%	94.1%	94.1%	11	9	93.5%	94.1%	94.1%	94.1%
RF	9	5	93.5%	94.1%	94.1%	94.1%	12	9	93.5%	94.1%	94.1%	94.1%
SVM	11	9	**96.7%**	96.9%	100%	94.1%	11	3	93.5%	94.4%	89.4%	100%

Table 3 – Average evaluation measures of k-fold cross-validation corresponding to the number of selected features and different classifiers implemented on the first dataset.

Classifier model	MI					RFE				
	No. of features	Accuracy	F1-score	precision	recall	No. of features	Accuracy	F1-score	precision	recall
LR	7	83.2%	85.5%	82.6%	89.4%	8	85.1%	87.2%	83.6%	91.7%
DT	3	85.1%	86.8%	84.7%	89.4%	11	79.3%	81.3%	80.6%	82.3%
RF	9	82.2%	84.1%	82.7%	85.8%	12	83.8%	85.7%	83.9%	88.2%
SVM	11	84.5%	86.7%	84%	90%	11	86.1%	88.3%	83.7%	94.1%

As is clear from the results in Table (2) that the mutual information method returns the significant information corresponding to each feature, while the number of selected features with high significant importance varies due to which classifier is used. Therefore, the proposed adaptive algorithm determines the appropriate number of features for each classifier that achieve higher classification accuracy.

The highest accuracy of about 96.7 is obtained using SVM with 11 features. Although the rest classifiers obtained higher accuracy with a smaller number of selected features the accuracy was less than the accuracy achieved by SVM. In medical applications, the importance of accuracy has a superior priority since it concerns patient life.

The reason behind these results is the impact of dataset size on the performance of a classifier. Where the DT and RF are affected by the size of the dataset more than SVM. So, the accuracy of RF decreases in a ratio higher than SVM as the size of the dataset decreases, due to the mechanism of RF that required a large dataset whereas, in SVM the

hyperplane depends on the support vector only so if the dataset has a support vector required, then the impact of dataset size will be irrelevant. These results come true with concluded outcomes by (Althnian et al., 2021).

As a comparison between MI and RFE, for all classifiers the classification accuracy that is achieved in the case of using RFE is less than the accuracy achieves using MI. so, for this database, the MI method is suitable more than RFE where the accuracy achieved by all classifiers using RFE was 93.5 which is obtained by using MI and DT with three selected features only.

Tables (4) and (5) show the number of selected features from the original second heart Statlog Cleveland Hungary database contains 1190 instances and 11 attributes, as well as the performance of each classifier used in terms of accuracy, f1-score, precision, and recall.

Table 4 – Maximum evaluation measures at specific k-fold cross-validation corresponding to the number of selected features and different classifiers implemented on the second dataset.

Classifier model	MI						RFE					
	No. of features	k-fold cross validation	Accuracy	F1-score	precision	recall	No. of features	k-fold cross validation	Accuracy	F1-score	precision	recall
LR	10	5	89.1%	89.4%	91.6%	87.3%	7	5	89.9%	90.1	93.2	87.3
DT	10	5	94.9%	95%	100%	90%	9	5	94.1%	94.1	100	88.8
RF	10	10	97.4%	97.6%	98.3%	96.8%	10	10	97.4%	97.6	98.3	96.8
SVM	9	6	93.2%	93.9%	89.8%	98.4%	6	5	91.5%	91.8	94.9	88.8

Table 5– Average evaluation measures of k-fold cross-validation corresponding to the number of selected features and different classifiers implemented on the second dataset.

Classifier model	MI					RFE				
	No. of features	Accuracy	F1-score	precision	recall	No. of features	Accuracy	F1-score	precision	recall
LR	10	81.7%	82.6	83.5	81.7	7	82.7%	83.6	84.2	83.1
DT	10	92.3%	92.7	93%	92.5	9	91.4%	91.8	93.1	90.6
RF	10	93.1	93.6	92.1	95.2	10	93.6%	94.1	93.3	94.9
SVM	9	90	91.2%	88.2%	94.4%	6	83.6%	84.5	84.5	83.8

Due to the increasing dataset size, the classification accuracy of DT and RF has improved as shown in Tables (4) and (5). Where the highest accuracy of 97.4% is obtained using RF with features selected using MI as well as RFE.

The comparison between the results in Table (3) and (4) show the impact of the increasing size of the database effect on the performance of RF and DT where the accuracy increases clearly. Whereas the performance of SVM and LR decreases (with a smaller number of selected features) due to the increasing of database dimension, where these classifiers are well done with high dimensionality. As is clear in Figures (2-5).

With this data set, there is no obvious difference in accuracy for all classifiers using features selected using MI or RFE, while the number of features selected using RFE is less than the number of features selected using MI with the same accuracy. That indicates that since the RFE mechanism is based on classifier criteria in calculating the weights to features then it is suitable for classifiers which not depend on information theory in calculating the decision such as LR and SVM.

All the above results can be visualized in Figure 2 - 5 to show the behavior of all classifier models across all features and all fold in the k-fold cross-validation method.

Figure 2 and Figure 3 describe the performance of four machine learning methods implemented on the first heart disease database in terms of f1-score with different numbers of features selected using MI and RFE-based feature selection methods and the proposed adaptive algorithm.

Fig. 2 - f1-score vs number of features different classifiers and MI method applying on first heart disease

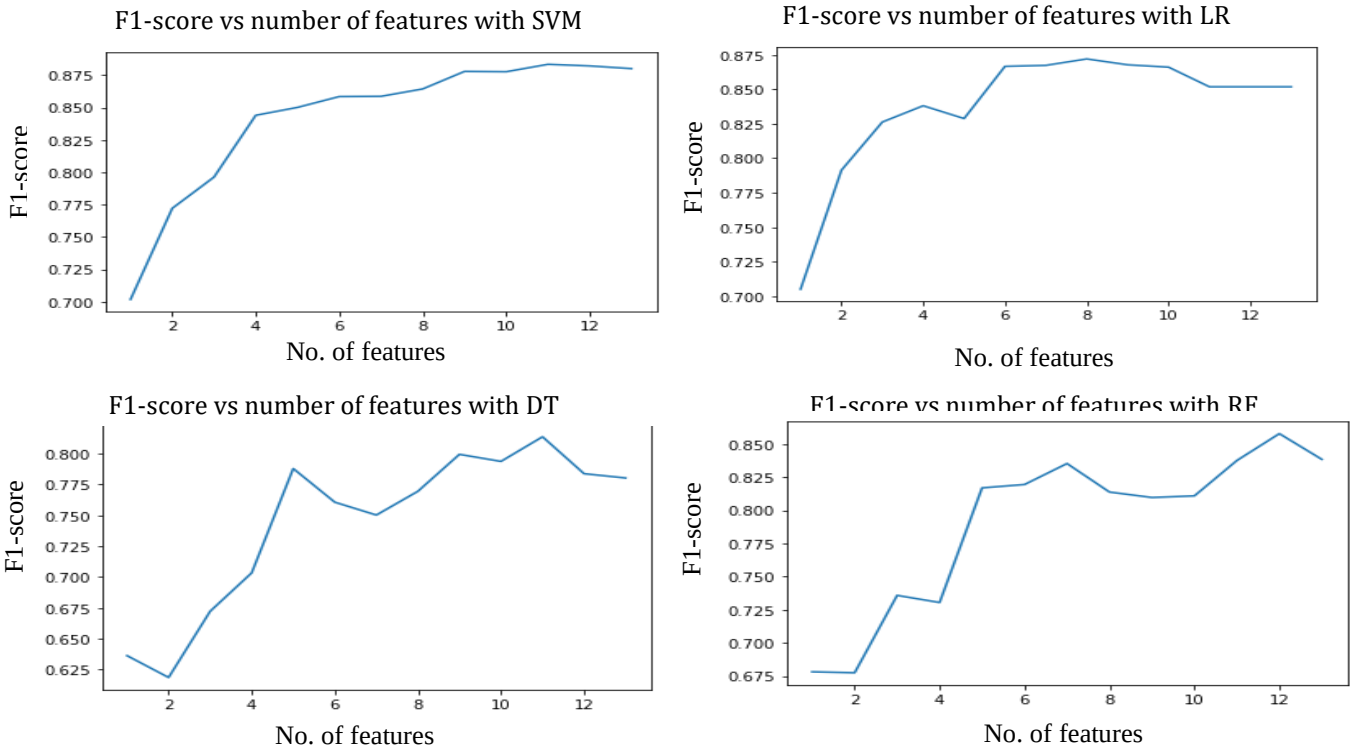

Fig. 3 - f1-score vs number of features different classifiers and RFE method applying on first heart disease database

Figure 4 and Figure 5 describe the performance of four machine learning methods implemented on the second heart disease database in terms of f1-score with the different number of features selected using MI and RFE-based feature selection methods and the proposed adaptive algorithm.

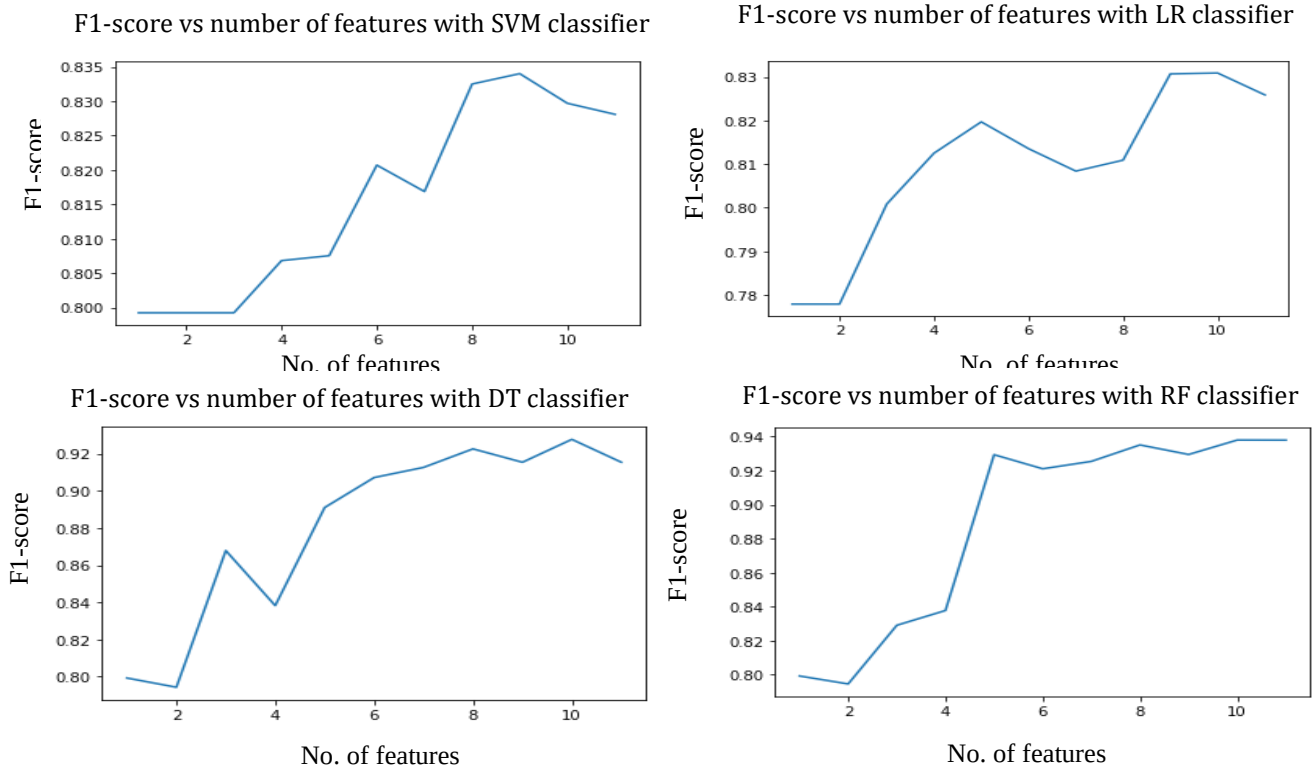

Fig. 4 - f1-score vs number of features different classifiers and MI method applying on second heart disease

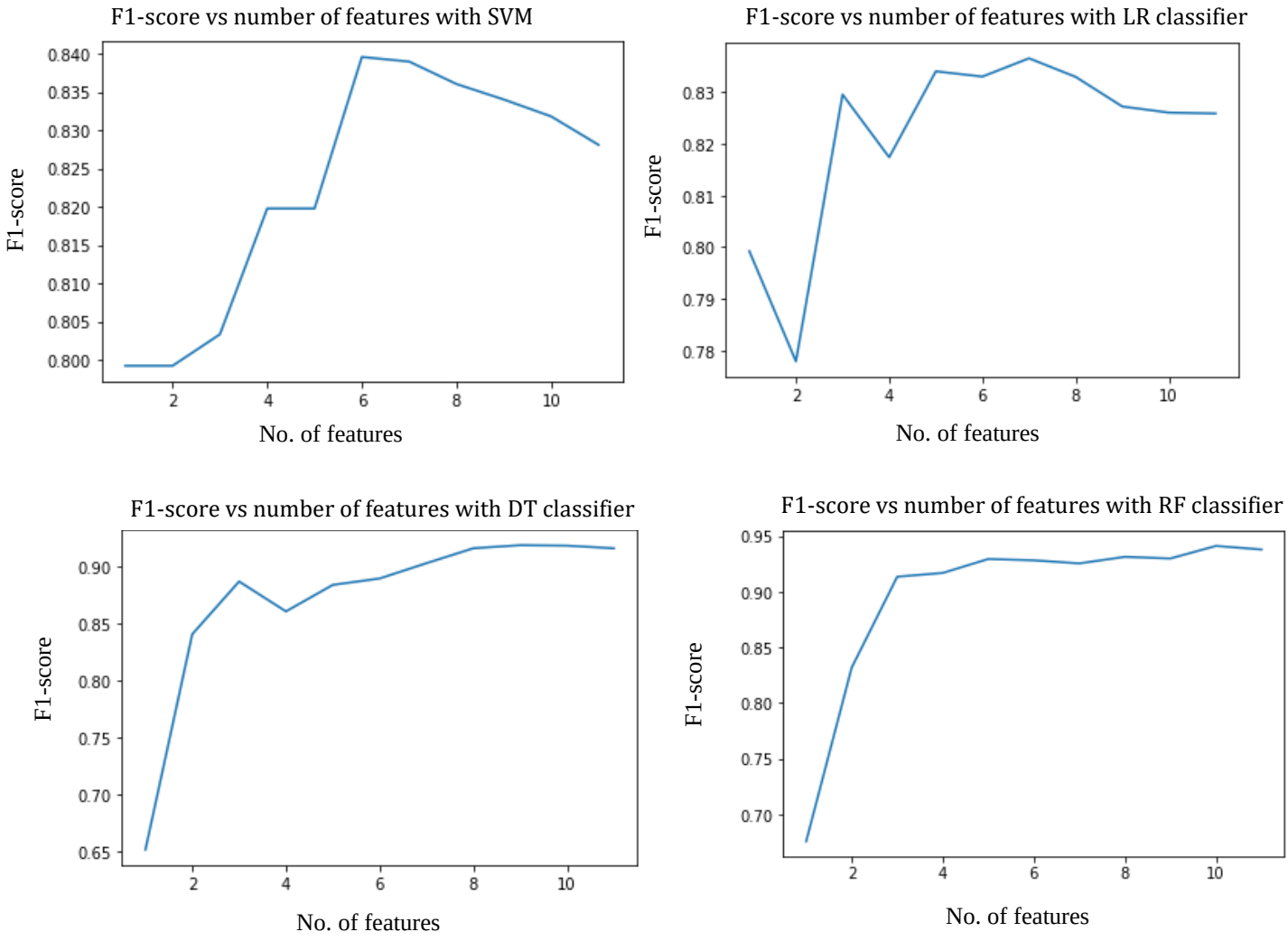

Fig.5: f1-score vs number of features different classifiers and RFE method applying on second heart disease database

As it is clear from the figures, each classification method required several selected features different from another classifier despite the mechanism of feature selection methods.

Although the MI-based feature selection method which is considered a filter-based feature selection method, works independently of the classifier method choosing the number of selected features depends on the classifier model so random thresholding way not always proper especially when the significant information for features is convergent, and when the classifier did well with the high dimension of the dataset.

In light of all the above results, the best model for each dataset will candidate as in Table (6) with a brief explanation of the reason for the choices for each model as the best model.

Table 6 – Candidate and the best model for each used dataset

Dataset	Candidate model	Feature selection method	Number of selected features	Accuracy	Reason of candidate	Best model
The first Cleveland heart disease	DT	MI	3	**93.5%**	Less number of selected features, acceptable accuracy	SVM-MI, and DT-MI, because MI has less complexity than RFE
	SVM	MI	11	**96.7%**	Highest accuracy	
	LR	RFE	8	93.5%	Less number of selected features, acceptable accuracy	
The heart Statlog Cleveland Hungary	DT	MI	10	**94.9%**	Less number of selected features, acceptable accuracy	DT-MI and RF-MI because MI has less complexity than RFE
	RF	MI	10	**97.4%**	Highest accuracy	
	DT	RFE	9	94.1%	Less number of selected features, acceptable accuracy	
	RF	RFE	10	97.4%	Highest accuracy	

The results in Table (6) show that the chosen best model is to be saved as a predictor model based on the highest accuracy, a smaller number of selected features, and the complexity of the feature selection method. At the first the priority for accuracy since the medical application is sensitive to accuracy more than other factors, then when there are two results with equality in classification accuracy the model with a smaller number of selected features is chosen. Finally, if two results are equal in accuracy and the number of selected features, the model with the less expensive feature selected model is chosen.

4.2 Comparison Study of proposed Classification Framework with Previous Works

The performance of the proposed technique is compared in terms of accuracy with previously existing classification techniques for heart disease. This comparison study illustrates in Table (7). Furthermore, the best candidate method for each dataset is compared and detected to be saved and incorporated into practical medical applications in health organizations.

Table 7 – Comparison of proposed technique performance and previous related works

References	Heart disease database	Feature selection method	Classifier model	Accuracy
[1]	Cleveland Heart Disease	fast conditional mutual information feature selection algorithm (FCMIM)	SVM	92.37
[9]	Cleveland Heart Disease	backward feature selection technique	DT	88.52%
[8]	Cleveland Heart Disease	Combination cumulative number	VOTE	88.41%
[11]	Cleveland Heart Disease	brute force method and possible Combination trials	(a hybrid technique with Naïve Bayes and Logistic Regression	87.4%
Our proposed technique	Cleveland Heart Disease	Mutual information	SVM	**96.7%**
	heart Statlog Cleveland Hungary database	Mutual information	RF	**97.4%**

5. Conclusion

In this work, an adaptive feature selection technique based on Mutual Information (MI) and Recursive Feature Elimination (RFE) methods has been produced to design a complete heart disease detection system. the proposed system design is implemented using four machine learning methods including Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF). The adaptive proposed algorithm aims to choose the optimal number of selected features instead of thresholding or randomly selecting. Two standard heart datasets are used to train the model from the UCI repository: the first Cleveland heart disease, and the second Statlog Cleveland Hungary. the results show that the highest accuracy is achieved by SVM with 11 features selected by mutual information-based feature selection method for the first database about 96.7, and by RF with 10 features selected by MI for the second database about 97.4%. the outcomes of this work indicate that for each dataset and each classifier as well as according to the used feature selection method, there is an optimal number of features in terms of classification accuracy. on the other hand, there is the best model candidate for each dataset in terms of a smaller number of features and acceptable classification accuracy to achieve the efficiency of the model has been saved. where the accuracy of 93.5% was achieved by DT with only 3 features selected by MI for the first dataset, and 93.2% achieved by SVM with only 9 features selected by MI for the second dataset. In the future, we aim to choose a model with a smaller number of features and improve it to achieve higher accuracy to ensure more efficiency required for fast and light prediction models that can be used in medical applications.

Acknowledgments

This study was completely supported by the authors and has not received any financial support from any organization.

References

[1] 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.

[2] P. Ghosh *et al.*, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques," *IEEE Access*, vol. 9, pp. 19304-19326, 2021.

[3] F. E. Harrell, "Ordinal logistic regression," in *Regression modeling strategies*: Springer, 2015, pp. 311-325.

[4] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine learning*, vol. 29, no. 2, pp. 131-163, 1997.

[5] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowledge and information systems*, vol. 14, no. 1, pp. 1-37, 2008.

[6] A. B. Shaik and S. Srinivasan, "A brief survey on random forest ensembles in classification model," in *International Conference on Innovative Computing and Communications*, 2019: Springer, pp. 253-260.

[7] D. Oreski, S. Oreski, and B. Klicek, "Effects of dataset characteristics on the performance of feature selection techniques," *Applied Soft Computing*, vol. 52, pp. 109-119, 2017.

[8] M. K. H. Eknath, "Identification of important characteristics and methods for data processing in cardiovascular estimation," *Journal of Emerging Technologies and Innovative Research*, vol. 8, no. 4, pp. 277-281, 2021. [Online]. Available: https://www.jetir.org/view?paper=JETIR2104038.

[9] K. Dissanayake and M. G. Md Johar, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2021, 2021.

[10] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, 2015: Ieee, pp. 1200-1205.

[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.

[12] A. Ul Haq, J. Li, M. H. Memon, J. Khan, and S. Ud Din, "A novel integrated diagnosis method for breast cancer detection," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 2, pp. 2383-2398, 2020.

[13] H. Zhou, X. Wang, and R. Zhu, "Feature selection based on mutual information with correlation coefficient," *Applied Intelligence*, vol. 52, no. 5, pp. 5457-5474, 2022.