International Journal on
"Technical and Physical Problems of Engineering"
(IJTPE)
Published by International Organization of IOTPE

ISSN 2077-3528
IJTPE Journal
www.iotpe.com
ijtpe@iotpe.com

December 2023 | Issue 57 | Volume 15 | Number 4 | Pages 152-159

UTILIZING PRINCIPAL COMPONENT ANALYSIS AND FEATURE SELECTION TO IMPROVE HEART DISEASE PREDICTION SYSTEM

A.F.H. Alharan [1] L.M. Saferali [1] S.K. Abbas [2] S.J. Mosa [1]

1. Computer Science Department, Faculty of Education for Girls, University of Kufa, Najaf, Iraq
abbasf.abood@uokufa.edu.iq, Luaymr@uokufa.edu.iq, safaaj@uokufa.edu.iq
2. College of Imam Kadhim for Islamic Science University, Baghdad, Iraq, elecnjf3@alkadhum-col.edu.iq

Abstract- One of the main causes of death known around the world is heart disease. Many systems and biomedical devices in hospitals contain vast amounts of clinical data. In order to increase prediction accuracy, it is vital to understand the facts about heart disease. In this research, a combination of principal component analysis (PCA) with Four wrapper feature selection methods, including Forward Feature Selection (FFS), Backward Feature Selection (BFS), Exhaustive Feature Selection (EFS), Recursive Feature Elimination (RFE), as well as four classifiers, including K-nearest Neighbor (KNN), Random Forest (RF), Decision Trees (DT) and XGBoost have been applied on two versions (let's say D1 and D2) of Cleveland heart disease datasets to analyse the results of hypothesis testing. With the KNN classifier, the data obtained by PCA-BFS, PCA-EFS, and PCA-RFE produced the maximum classification accuracy of 95.08% for D1. For D2, all classifiers had the highest accuracy of 100% when using PCA with all feature selection methods.

Keywords: Heart Disease, Feature Selection, Feature Extraction, Machine Learning, PCA.

1. INTRODUCTION

Heart disease HD is a significant public health concern, and many people around the world have died as a result of it. Shortness of breath, swollen feet, and general physical weakness are all symptoms commonly associated with HD [1]. Current methods of diagnosing HD aren't adequate for early detection for a number of reasons, such as how accurate they are and how long they take to do. As a result, researchers are working to develop an effective method for the early detection of HD [2]. It is well known that HD is difficult to diagnose and treat in the absence of modern technology and when medical specialists are not available [1]. Early detection and treatment can help save the lives of many people [3].

The European Society of Cardiology estimates that 26 million people worldwide have HD, with 3.6 million new cases diagnosed annually [4]. The vast majority of individuals have HD [2]. Traditionally, HD has been diagnosed using a physician's analysis of the medical history of the patient, analysis of concerned symptoms, and physical examination reports. The results of this diagnostic technique, however, do not accurately identify HD patients. In addition, physicians frequently use electrocardiograms (ECGs), stress testing (stress ECGs, exercise stress tests, and nuclear cardiac stress tests), cardiac Magnetic Resonance Imaging (MRIs), echocardiograms (heart ultrasounds), and angiography to diagnose cardiovascular conditions. As a consequence, the general public cannot afford the costs associated with cardiovascular disease diagnosis and treatment. Data mining technologies predict the rapid and accurate identification of patients who have an increased risk of developing HD, thereby reducing diagnostic and treatment costs [5]. Many researchers have examined feature selection methods and multiple classifiers using two versions of Cleveland HD datasets [6-11].

Existing research has proposed a variety of diagnosis techniques based on machine learning as a way to diagnose HD. In order to increase prediction accuracy, this research study proposed and developed a new framework of machine learning-based diagnosis techniques. Two versions of Cleveland HD datasets are used to test the suggested method in this article. In the beginning, the two datasets are pre-processed and cleaned, as referred to in section 2. Further, four types of experiments are conducted to analyze the data in pre-processing. Initially, the pre-processed data evaluated all four classifiers. The second experiment involves the utilization of PCA to extract features from the pre-processed data. These extracted features are subsequently used as input for the classifiers. The final experiment employs four common wrapper-based feature selection methods (FFS, BFS, EFS, and RFE) to obtain a reduced set of pre-processed datasets. Then, feature extraction using PCA is conducted on the reduced datasets, which are then validated with the classifiers. The accuracy metric is utilized to measure the performance of the presented module. The framework of this approach is shown in Figure 1.

The remaining sections of the document have been structured in the following manner: Section 2 will review the literature on HD prediction systems, followed by a comprehensive explanation of the research methodology presented in Section 3. The experimental results and their

corresponding discussion are presented in Section 4, while Section 5 will present the conclusions and outline future research directions.

2. LITERATURE REVIEW

Aggrawal and Saurabh [12] introduced a new sequential feature selection technique to recognize deaths that occurrences in HD patients while receiving medical treatment, in order to determine the important features. Many machine learning techniques are employed such as RF, DT, Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Gradient Boosting (GBC), and KNN. In order to validate the results of the Sequential Forward Selection (SFS) algorithm many parameters are generated including (precision, confusion matrix, F1-score, receiver operating characteristic curve, and recall rate). The results showed that (sequential feature selection) by using the RF classifier method achieved the highest accuracy of 86.67%.

Takci [13] predicted heart attacks using 12 classification algorithms from various categories in addition to 4 feature selection methods. The models evaluated depend on (the results of ROC analysis, processing time, and accuracy) Without using any type of feature selection, the highest accuracy was 82.59%, while with using feature selection techniques, the model accuracy achieved to 84.81% based on (linear SVM and naive Bayes). A reduction in processing time from (359 milliseconds) to (187 milliseconds) was also occurred. Based on the mean accuracy value, between four the various alternative feature selection methods using ReliefF algorithm produces the highest accuracy. Therefore, the author concluded that using the appropriate feature selection is useful for HD prediction methods.

Xiao Yan, et al. [9] proposed a prediction model for HD, where they combined ensemble methods (boosting and bagging) in addition the author used two feature extraction algorithms LDA and PCA in addition to using five classifier algorithms including (KNN, SVM, RF, NB, and DT) were evaluated on Cleveland HD dataset subsets and compared to ensemble methods (bagging and boosting). According to the findings of the experiments, the best performance was achieved by using the bagging ensemble learning technique combined with two feature extraction algorithms PCA and DT. Latha and Jeeva [14] used ensemble classification and feature selection to predict HD risk. Ensemble methods such as bagging and boosting improve weak classifier prediction accuracy and HD risk prediction. Weak classifiers improved by 7% with ensemble classification. Adding feature selection improved prediction accuracy and performance. Spencer [10] tested Chi-squared algorithm, PCA, symmetrical uncertainty, and ReliefF on four common HD datasets. The authors noticed that feature selection benefits vary by cardiac dataset machine learning approach. Chi-squared feature selection and the Bayes-Net classifier produced the most accurate models, with 85.0% accuracy, 84.73% precision, and 85.56% recall.

Almansour [15] compared two classifiers, SVM algorithm and ANN algorithm, and used a random exhaustive search method to improve their parameters in order to help in chronic kidney disease early detection. The features were selected using correlation coefficients after pre-processing of 400 instances dataset from UCI repository. The effectiveness of classifiers is evaluated in relation to the number of training cycles needed and the best features (namely F2, F3, F6, and all). The 12 best features were subsequently utilized to forecast renal illness by using (SVM algorithm and ANN algorithm) because ANN algorithm is more accurate than SVM algorithm. Literature shows that train the classifiers with appropriate features that selected by using a different feature selection algorithm can increase classifier accuracy.

3. FRAMEWORK AND METHODOLOGY

The primary goal of the study is to enhance the accuracy of classification by extracting the feature of the reduced feature of a dataset related to cardiac disease.

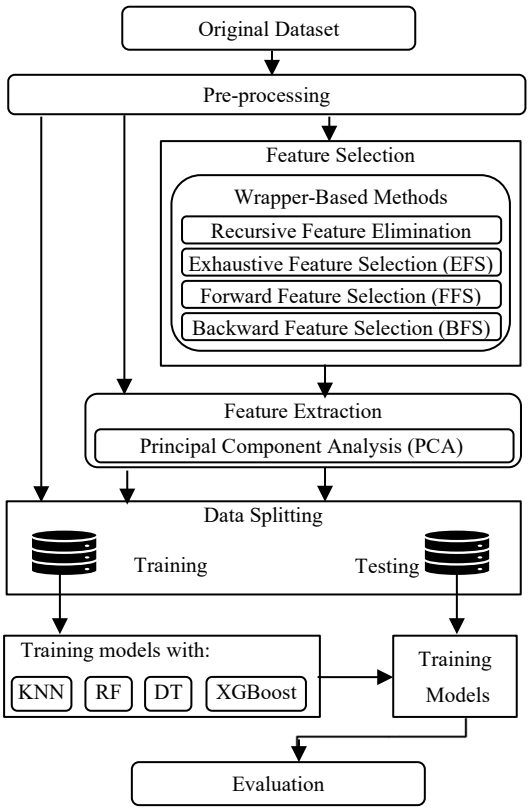

Figure 1. The proposed framework

Figure 1 describes the structure employed to categorize heart-related diseases, which encompasses various components such as data gathering, data pre-processing, feature selection, feature extraction, data splitting, Classifier-based training of the model and subsequent evaluation of its performance. The subsequent sections delineate the fundamental elements of the proposed framework as illustrated in Figure 1.

3.1. Data Collection

For the purpose of testing, the study utilized two versions of the Cleveland HD datasets obtained from UCI repository and Kaggle repository [16], which is accessible

online. The first D1 consisted of 303 instances and 14 features as it is used in the published studies [6], [12]. The second D2 is a set of same the 14 features D1 but with 1025 instances [9], [17]. The result feature is divided into two categories that signify the presence or absence of HD. The D1 includes 138 cases as healthy people, and 165 have the HD. While in D2, 499 are healthy and 526 have the disease. Table 1 provides the description of HD dataset [6].

Table 1. Summary of dataset

Feature ID	Features	Description
F1	Age	Patient's age
F2	Sex	1=male, 0=female
F3	Chest pain (CP)	CP type (1=typical angina, 2=atypical angina, 3=nonanginal, 4=asymptomatic)
F4	RestBP	Resting blood pressure
F5	Chol	Serum cholesterol in mg/dl
F6	FBS	Fasting blood sugar larger 120 mg/dl (1 true)
F7	RestECG	Resting electrocardiographic result
F8	Thalach	Maximum heart rate obtained
F9	Exang	Exercise-induce angina (1 yes)
F10	Oldpeak	ST depression induce: exercise relative to rest
F11	CA	Number of major vessels (0-3)
F12	Slope	Slope of peak exercise ST
F13	Thal	Heart status (7=reversible defect, 6=fixed defect, 3=normal)
F14	Num	Diagnosis of heart disease (1=yes, 0=No)

3.2. Data Pre-Processing

To ensure that data quality is accurately represented, it is necessary to pre-process the dataset. In this study, the data is scaled by StandardScaler and MinMaxScaler. As well as the missing value replaced by the mean value of feature.

3.3. Feature Selection

In the machine learning paradigm, feature selection is a crucial pre-processing step that identify the optimal subset of features by eliminating irrelevant and redundant information [18]. The present study has implemented various feature selection techniques to choose a different number of features (NF) which are 7, 8, 9 and 10. Specifically, four distinct feature selection techniques have been utilized, as explained below:

In this study, Different number of features (NF) is selected using different techniques. This study has employed four feature selection techniques, which are outlined below:

3.3.1. Forward Feature Selection (FFS)

FFS is a feature selection technique used in machine learning to identify the most important features or variables for a given predictive model. It is a computationally efficient method and can be used for both regression and classification tasks. The steps of the (FFS) algorithm can be summarized as follows:

1. Initialize an empty set of selected features.
2. Evaluate the performance of the model with each individual feature added to the selected set.
3. Select the feature that results in the highest improvement in model performance.
4. Add the selected feature to the set of selected features.
5. Repeat steps 2-4 until including *NF* features.

3.3.2. Backward Feature Selection (BFS):

The BFS strategy is the total antithesis of the FFS strategy [12]. The BFS procedure is completed by the steps listed below.

1. Initialize the set of selected features to be the full set of features.
2. Evaluate the performance of the model with each individual feature removed from the selected set.
3. Select the feature whose removal results in the least decrease in model performance.
4. Remove the selected feature from the set of selected features.
5. Repeat steps 2-4 until including the NF features.

3.3.3. Exhaustive Feature Selection (EFS):

It is a brute force technique used in machine learning to identify the best subset of features that result in the highest model performance [19]. The steps of EFS are listed as the following:

1. Generate all possible feature combinations with NF features from the full set of features.
2. For each feature combination, train a model using only the selected features and evaluate its performance using an appropriate metric, such as accuracy or mean squared error.
3. Select the feature combination that results in the highest model performance.
4. Repeat steps 1-3 for different NF features.

3.3.4. Recursive Feature Elimination (RFE)

This technique for selecting features eliminates them from the dataset in a recursive manner, ultimately choosing the optimal subset of features that produces the most effective model performance [20]. The algorithm starts with all set of features then removes the least important features one at a time until reached the desired number of features. The steps of the (RFE) algorithm can be summarized as follows:

1. Train a model using all available features in the dataset.
2. Evaluate the importance of each feature in the model using a specified metric, such as coefficients of a linear model or feature importance's of a tree-based model.
3. Remove the least important feature(s) from the dataset.
4. Repeat steps 1-3 until *NF* features are reached.

3.4. Feature Extraction

Selecting the most appropriate features is a vital step since irrelevant ones can negatively impact the machine learning classifier's classification efficiency. PCA [21], [22] is employed during this phase to identify and choose the essential features from the dataset.

3.5. Data Splitting

In this study, the dataset is distributed into two parts: 80% and 20%. The larger portion, 80%, is designated as the training part of dataset, while the part 20% is used as the testing dataset. The training part of dataset is used to train a model, while the testing part of dataset is utilized to assess the model's performance.

3.6. Classification Techniques

The utilization of machine learning techniques involves the prediction of a problem's output by teaching the classifier through historical data that has been labelled. This study has employed four commonly used machine learning techniques. The following is a description of each classification technique:

3.6.1. K-Nearest Neighbor (KNN)

The KNN technique is a straightforward machine learning algorithm employed in both classification and regression operations [23]. The function of this algorithm entails measuring the distance that exists between a new data point and every pre-existing data points within the dataset. The *K* closest points are then considered for classification or regression. K-NN is a non-parametric algorithm and easy to implement, but it can be computationally expensive and sensitive to distance metric choice and *K* value [24]. for more information about K-Nearest Neighbors (KNN), refer to [24].

3.6.2. Random Forest (RF)

RF is an ensemble classification technique that is widely used and well-developed. This method generates multiple decision trees by utilizing different subsets of data during the training phase. In the testing phase, each decision tree assigns a class label to the corresponding data point. The final outcome is then calculated based on the majority vote of all the decision trees. By comparing the number of votes for each class, we can determine the correct label and significantly enhance the accuracy of prediction [25].

3.6.3. Decision Tree (DT)

It is commonly used algorithm in the field of machine learning that represents the outcomes of various decisions. [26]. It splits the data into subsets based on input feature values using a metric like information gain. The tree is constructed by recursively selecting the best feature and split. The prediction is made by traversing the tree from root to leaf node. The algorithm chooses the branch that matches the value of the input feature. The splitting criterion can be represented mathematically using the entropy formula H(S) and the gain formula Gain. For further information about decision trees, you can refer to the provided reference [26].

3.6.4. XGBoost

XGBoost is an optimized (GBC) machine learning algorithm [27]. It builds an ensemble of (DT) using a combination of additive and tree-pruning methods. The objective function to optimize is a sum of loss function and regularization term. The prediction is made by adding up the predictions from all the trees. The algorithm can handle missing values and monotonic constraints. It is popular in winning machine learning competitions. The objective function can be represented mathematically using the Taylor series approximation. For additional details regarding XGBoost, you can find more information in [27].

3.7. Model Evaluation

The proposed model is assessed using accuracy criteria. In classification tasks, accuracy is regarded as a crucial performance metric. The accuracy is calculated by dividing the number of samples by the number of correct classifications, as demonstrated in Equation (1) [28].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Four essential indicators are utilized to assess a model's performance in binary classification. *TP* is the count of instances appropriately categorized as positive, whereas *TN* is the count of instances appropriately categorized as negative. *FP* is the count of instances inappropriately categorized as positive, whereas *FN* is the count of instances inappropriately categorized as negative.

4. RESULTS AND DISCUSSION

The proposed framework in this study was implemented using Python programming language in the Jupyter (Anaconda) Notebook platform, which facilitated the identification of patterns and efficient exploration of the dataset. Additionally, Python provides a wide range of libraries and tools that are suitable for feature selection, feature extraction, and classification tasks in machine learning and data analysis, among which are the widely used scikit-feature and scikit-learn libraries. As mentioned in section 3.1., D1 and D2 are used for analysis. After data pre-processing, four different feature selection methods were utilized to choose the best possible subsets of features, namely, FFS, BFS, EFS, RFE. The objective was to identify the highest-ranking feature subsets.

Following the feature selection process, the chosen data was subjected to feature extraction using PCA. The purpose of using PCA is to identify the most important components of the selected features while retaining as much information as possible. This is done to improve the efficiency and effectiveness of the subsequent analysis and modelling steps. Table 2 displays the chosen attributes for the D1 dataset utilizing FFS, BFS, EFS, and RFE techniques, while Table 3 demonstrates the selected features for the D2 dataset utilizing the same methods. Each method is conducted with subsets containing a varying number of features, namely 7, 8, 9, and 10. The attributes in the tables are represented by either 1 or 0, where 1 indicates that the attribute is included in the feature subset and 0 indicates that it is not included.

Based on the Table 2, the features that hold the highest significance for predicting HD are (F3, F9, F10, F12, and F13). The other features are following in order of occurrence. However, the ranking of these features varies depending on the method used for feature selection. According to Table 3, (F2, F3, F8, F10, F12, and F13) are the most important features. The other features follow a sequential order based on their appearance. In this study, PCA was used as feature extraction method to transform the data subsets into manageable and simplified form that retains as much useful information as possible. After applying feature selection techniques and PCA to the data, the resulting data was divided into two sets: a training set, which comprises 80% of the data and will be used to train machine learning models, and a testing set, which consists of the remaining 20%.

Table 2. Selected features from D1 dataset using FFS, BFS, EFS and RFE methods

Feature selection method	Features													Number of features
	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	
(FFS)	1	1	1	0	1	1	0	1	1	1	0	1	1	10
	0	1	1	0	1	1	0	1	1	1	0	1	1	9
	0	0	1	0	1	1	0	1	1	1	0	1	1	8
	0	0	1	0	1	1	0	0	1	1	0	1	1	7
(BFS)	0	1	1	1	0	0	1	1	1	1	1	1	1	10
	0	1	1	1	0	0	1	1	1	1	0	1	1	9
	0	1	1	0	0	0	1	1	1	1	0	1	1	8
	0	1	1	0	0	0	0	1	1	1	0	1	1	7
(EFS)	0	1	1	0	0	1	1	1	1	1	1	1	1	10
	0	1	1	0	0	1	1	1	1	1	0	1	1	9
	0	0	1	0	0	0	1	1	1	1	1	1	1	8
	1	0	1	0	0	0	1	0	1	1	0	1	1	7
(RFE)	0	1	1	1	0	1	1	1	1	1	0	1	1	10
	0	1	1	1	0	0	1	1	1	1	0	1	1	9
	0	1	1	0	0	0	1	1	1	1	0	1	1	8
	0	1	1	0	0	0	0	1	1	1	0	1	1	7
Total number of occurrences	2	12	16	4	4	7	10	14	16	16	3	16	16	

Table 3. Selected features from D2 dataset using FFS, BFS, EFS and RFE methods

Feature selection method	Features													Number of features
	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	
FFS	1	1	1	1	0	0	1	1	0	1	1	1	1	10
	0	1	1	1	0	0	1	1	0	1	1	1	1	9
	0	1	1	1	0	0	1	1	0	1	0	1	1	8
	0	1	1	1	0	0	0	1	0	1	0	1	1	7
BFS	0	1	1	1	1	0	0	1	1	1	1	1	1	10
	0	1	1	1	0	0	0	1	1	1	1	1	1	9
	0	1	1	1	0	0	0	1	1	1	0	1	1	8
	0	1	1	0	0	0	0	1	1	1	0	1	1	7
EFS	1	1	1	1	0	0	1	1	0	1	1	1	1	10
	0	1	1	0	0	0	1	1	1	1	1	1	1	9
	0	1	1	0	0	0	1	1	0	1	1	1	1	8
	0	1	1	0	0	0	1	1	0	1	0	1	1	7
RFE	0	1	1	1	1	0	0	1	1	1	1	1	1	10
	0	1	1	1	0	0	0	1	1	1	1	1	1	9
	0	1	1	0	0	0	0	1	1	1	1	1	1	8
	0	1	1	0	0	0	0	1	1	1	0	1	1	7

Table 4. Classification accuracy of combination PCA and the feature section techniques (FFS, BFS, EFS and RFE) (D1 dataset)

Feature selection technique	With PCA	Number of features	*KNN*	*RF*	*DT*	XGBoost
Original dataset	No	13	91.80	88.52	80.33	81.97
Original dataset (PCA)	Yes	13	91.80	81.97	75.41	81.97
FFS	Yes	10	93.44	88.52	81.97	80.33
		9	93.44	90.16	83.61	85.25
		8	91.80	88.52	80.33	85.25
		7	91.80	81.97	70.49	78.69
(BFS)	Yes	10	90.16	88.52	81.97	83.61
		9	91.80	90.16	81.97	86.89
		8	91.80	90.16	88.52	91.16
		7	95.08	90.16	80.33	88.52
(EFS)	Yes	10	91.80	86.89	86.89	86.89
		9	95.08	90.16	78.69	85.25
		8	90.16	83.61	77.05	83.61
		7	90.16	83.61	80.33	81.97
(RFE)	Yes	10	93.44	90.16	81.97	88.52
		9	91.80	90.16	81.97	86.89
		8	91.80	90.16	88.52	91.80
		7	95.08	90.16	80.33	88.52

Table 4 and Table 5 show the accuracies of predicting HD for D1 and D2 datasets respectively. The results obtained by using the combination of most effective feature set (FFS, BFS, EFS and RFE) and PCA across multiple classification algorithms. As well as, Figure 2 and Figure 3 provided an overview of the results in Table 4 and Table 5, respectively.

Table 5. Classification accuracy of combination PCA and the feature section techniques (FFS, BFS, EFS and RFE) (D2 dataset)

Feature selection technique	With PCA	Number of features	*KNN*	*RF*	*DT*	XGBoost
Original dataset	No	13	98.54	98.54	98.54	98.54
Original dataset (PCA)	Yes	13	98.54	98.54	98.54	98.54
(FFS)	Yes	10	98.54	100	100	100
		9	98.54	98.54	97.07	100
		8	98.54	100	97.07	98.54
		7	100	98.54	98.54	98.54
BFS	Yes	10	98.54	98.54	98.54	100
		9	100	100	100	100
		8	100	98.54	98.54	97.07
		7	100	100	100	100
EFS	Yes	10	98.54	100	100	100
		9	100	98.54	97.07	98.54
		8	100	98.54	97.07	100
		7	98.54	97.07	98.54	98.54
RFE	Yes	10	98.54	98.54	98.54	100
		9	100	100	100	100
		8	100	100	98.54	100
		7	100	100	100	100

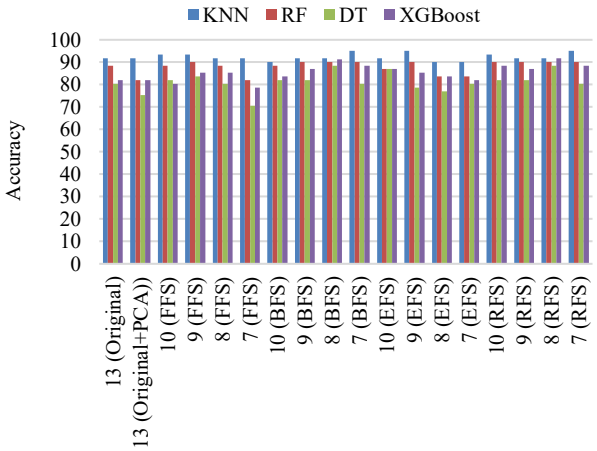

Figure 2. Classification accuracies of FFS-PCA, BFS-PCA, EFS-PCA and RFE-PCA (D1 dataset)

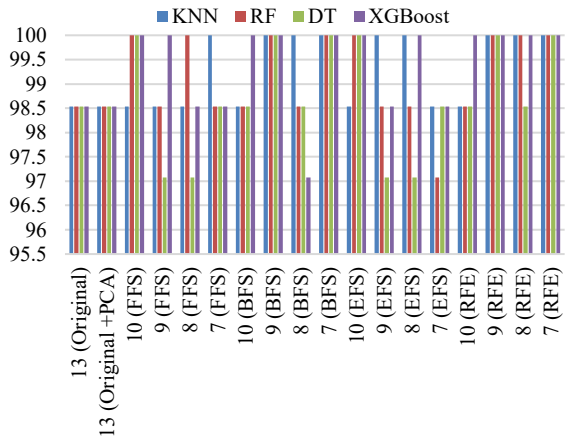

Figure 3. Classification accuracies of FFS-PCA, BFS-PCA, EFS-PCA and RFE-PCA (D2 dataset)

As Table 4, the original data has achieved 91.80% as a maximum accuracy using KNN, while the original data with PCA has achieved same the accuracy by KNN also. For FFS-PCA, the best accuracy is 93.55% that achieved by KNN when the number of features is 9 (namely 2, 3, 5, 6, 8, 9, 10, 12, 13) and 10 (namely 1, 2, 3, 5, 6, 8, 9, 10, 12, 13). The maximum accuracy achieved for D1 dataset is 95.08% using KNN when BFS-PCA, EFS-PCA and RFE-PCA are applied, where the number of selected features by BFS, EFS and RFE are 7 (namely 2, 3, 8, 9, 10, 12, 13), 9 (namely 2, 3, 6, 7, 8, 9, 10, 12, 13) and 7 (namely 2, 3, 8, 9, 10, 12, 13), respectively.

For the results in Table 5, the original data with and without PCA have achieved 98.54% as a maximum accuracy using KNN. Combination PCA with All feature selection techniques have achieved accuracy 100% using all classifiers and for different number of selected features. The summary of maximum accuracy with D1 and D2 datasets is illustrated in Table 6 and Table 7, respectively.

Table 8 presents a comparison between the proposed work and other previously published works. The table demonstrates that the proposed work achieved a higher accuracy compared to the others, particularly for the D1 and D2 datasets. As a result, it can be inferred that the proposed model significantly outperforms its competitors.

Table 6. Summary of maximum accuracy (D1)

Feature selection method	Maximum Accuracy	Classifier	Number of features
Original dataset (without PCA)	91.80	KNN	13
Original dataset (with PCA)	91.80	KNN	13
FFS+PCA	93.44	KNN	9,10
BFS+PCA	95.08	KNN	7
EFS+PCA	95.08	KNN	9
RFE+PCA	95.08	KNN	7

Table 7. Summary of maximum accuracy (D2)

Feature selection method	Maximum Accuracy	Classifier	Number of features
Original dataset (without PCA)	98.54	KNN, RF, DT, XGBoost	13
Original dataset (with PCA)	98.54	KNN, RF, DT, XGBoost	13
FFS+PCA	100	KNN	7
FFS+PCA	100	RF	8, 10
FFS+PCA	100	DT	10
FFS+PCA	100	XGBoost	9, 10
BFS+PCA	100	KNN	7, 8, 9
BFS+PCA	100	RF	7, 9
BFS+PCA	100	DT	7, 9
BFS+PCA	100	XGBoost	7, 9, 10
EFS+PCA	100	KNN	8, 9
EFS+PCA	100	RF	10
EFS+PCA	100	DT	10
EFS+PCA	100	XGBoost	8, 10
RFE+PCA	100	KNN	7, 8, 9
RFE+PCA	100	RF	7, 8, 9
RFE+PCA	100	DT	7, 9
RFE+PCA	100	XGBoost	7, 8, 9, 10

Table 8. Comparison between the result of the proposed work and other existed works

Method	Reference	Accuracy
D1 Dataset		
Information Gain +SVM Chi-square +SVM	[21] 2021	83.41%
BFS+DT	[11] 2021	88.52%
RFFS+RF	[19] 2021	85.25%
XGB	[17] 2022	91.6%
FFS+PCA +KNN	Proposed Method	93.44%
BFS+PCA+KNN		95.08%
EFS+PCA+KNN		
RFE+PCA+KNN		
D2 dataset		
PCA+DT	[9] 2021	98.6%
Ada-Boost	[29] 2020	97%%
KNN	[30] 2020	99.71%
(FFS+PCA, BFS+PCA, EFS+PCA, RFE+PCA) with all classifiers (KNN, RF, DT, XGBoost)	The Proposed Method	100%

5. CONCLUSION AND RUTURE WORKS

The main objective of this study is to investigate how feature selection and feature extraction methods impact the accuracy of predicting heart disease. To achieve this objective, an evaluation was carried out on a set of essential features obtained from the commonly employed Cleveland HD datasets, which are accessible at UCI. Four wrapper-based feature selection methods were utilized for this analysis. PCA was utilized to extract the features from the subset of important features. Four classifiers were implemented to predict the heart disease. By using D1 dataset, the highest accuracy obtained by KNN was 95.05% when PCA used with BFS, EFS and RFE for subset of 7, 9, 7 features out of 13 features, respectively.

For D2 dataset, the classifiers KNN, RF, DT and XGBoost was obtained 100% as the highest accuracy with all feature selection method for different data subsets of 7, 8, 9 and 10 features. Form the result, the suggested method outperformed the other methods in published studies in terms of accuracy. As a future works, we are exploring the use of deep learning methods for heart disease prediction. Furthermore, integrating multiple resources or other types of data can providing comprehensive understanding of heart disease and improve the accuracy of prediction models.

NOMENCLATURES

Acronyms

HD	Heart disease
PCA	Principal Component Analysis
FFS	Forward Feature Selection
BFS	Backward Feature Selection
EFS	Exhaustive Feature Selection
RFE	Recursive Feature Selection
KNN	K-Nearest Neighbor
RF	Random Forest
DT	Decision Tree

REFERENCES

[1] A.L. Bui, T.B. Horwich, G.C. Fonarow, "Epidemiology and Risk Profile of Heart Failure", Nat. Rev. Cardiol., Vol. 8, No. 1, pp. 30-41, 2011.

[2] L.A. Allen, et al., "Decision Making in Advanced Heart Failure: A Scientific Statement from the American Heart Association", Circulation, Vol. 125, No. 15, pp. 1928-1952, 2012.

[3] World Health Organization, "Cardiovascular Diseases", WHO, Geneva, Switzerland, 2020. www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1.

[4] J. Lopez Sendon, "The Heart Failure Epidemic", Medicographia, Vol. 33, No. 4, pp. 363-369, 2011.

[5] K. Vanisree, J. Singaraju, "Decision Support System for Congenital Heart Disease Diagnosis Based on Signs and Symptoms Using Neural Networks", Int. J. Comput. Appl., Vol. 19, No. 6, pp. 6-12, April 2011.

[6] S.I. Ayon, M.M. Islam, M.R. Hossain, "Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques", IETE J. Res., Vol. 68, No. 4, pp. 2488-2507, 2022.

[7] K. Srivastava, D.K. Choubey, "Heart Disease Prediction using Machine Learning and Data Mining", Int. J. Recent Technol. Eng., Vol. 9, No. 1, pp. 212-219, 2020.

[8] J.P. Li, A.U. Haq, S.U. Din, J. Khan, A. Khan, A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare", IEEE Access, Vol. 8, No. M l, pp. 107562-107582, 2020.

[9] X.Y. Gao, A. Amin Ali, H. Shaban Hassan, E.M. Anwar, "Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method", Complexity, Vol. 2021, pp. 1-10, February 2021.

[10] R. Spencer, F. Thabtah, N. Abdelhamid, M. Thompson, "Exploring Feature Selection and Classification Methods for Predicting Heart Disease", Digit. Heal., Vol. 6, pp. 1-10, 2020.

[11] K. Dissanayake, G. Johar, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms", Hindawi Appl. Comput. Intell. Soft Comput., Vol. 2021, p. 17, 2021.

[12] R. Aggrawal, S. Pal, "Sequential Feature Selection and Machine Learning Algorithm-Based Patient's Death Events Prediction and Diagnosis in Heart Disease", SN Comput. Sci., Vol. 1, No. 6, pp. 1-16, 2020.

[13] H. Takci, "Improvement of Heart Attack Prediction by the Feature Selection Methods", Turkish J. Electr. Eng. Comput. Sci., Vol. 26, No. 1, pp. 1-10, 2018.

[14] C.B.C. Latha, S.C. Jeeva, "Improving the Accuracy of Prediction of Heart Disease Risk Based on Ensemble Classification Techniques", Informatics Med. Unlocked, Vol. 16, No. June, p. 100203, June 2019.

[15] N.A. Almansour, et al., "Neural Network and Support Vector Machine for the Prediction of Chronic Kidney Disease: A Comparative Study", Comput. Biol. Med., Vol. 109, No. October 2018, pp. 101-111, 2019.

[16] Heart Disease, "UCI Machine Learning Repository", https://archive.ics.uci.edu/ml/datasets/Heart+Dise

[17] E.M. Abd Allah, D.E. El Matary, E.M. Eid, A.S.T. El Dien, "Performance Comparison of Various Machine Learning Approaches to Identify the Best One in Predicting Heart Disease", J. Comput. Commun., Vol. 10, No. 2, pp. 1-18, 2022.

[18] J. Cai, J. Luo, S. Wang, S. Yang, "Feature Selection in Machine Learning: A New Perspective", Neurocomputing, Vol. 300, pp. 70-79, 2018.

[19] A.A. Abdullah, N.A. Alhadi, W. Khairunizam, "Diagnosis of Heart Disease Using Machine Learning Methods", Intelligent Manufacturing and Mechatronics: Proceedings of Sympo SIMM 2020, Springer, pp. 77-89, 2021.

[20] E.M. Senan, et al., "Diagnosis of Chronic Kidney Disease using Effective Classification Algorithms and Recursive Feature Elimination Techniques", J. Healthc. Eng., Vol. 2021, 2021.

[21] P. Khurana, S. Sharma, A. Goyal, "Heart Disease Diagnosis: Performance Evaluation of Supervised Machine Learning and Feature Selection Techniques", The 8th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 510-515, 2021.

[22] A.K. Garate Escamila, A.H. El Hassani, E. Andres, "Classification Models for Heart Disease Prediction Using Feature Selection and PCA", Informatics Med. Unlocked, Vol. 19, p. 100330, 2020.

[23] V. Yousefi, S. Kheiri, S. Rajebi, "Evaluation of K-Nearest Neighbor, Bayesian, Perceptron, RBF and SVM Neural Networks in Diagnosis of Dermatology Disease", International Journal on Technical and Physical Problem on Engineering (IJTPE), Issue 42, Vol. 12, No. 1, pp. 114-120, March 2020.

[24] E. Alpaydin, "Introduction to Machine Learning", MIT Press, 2020.

[25] A. Cutler, "Random Forests for Regression and Classification", Utah State Univ. Ovronnaz, Switz., Switzerland, 2010.

[26] J. Han, J. Pei, H. Tong, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2022.
[27] T. Chen, C. Guestrin, "Xgboost: A Scalable Tree Boosting System", The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, August 2016.
[28] S. Kumar, H. Kumar, "LUNGCOV: A Diagnostic Framework Using Machine Learning and Imaging Modality", International Journal on Technical and Physical Problem on Engineering (IJTPE), Issue 51, Vol. 14, No. 2, pp. 190-199, June 2022.
[29] G. Choudhary, S.N. Singh, "Prediction of Heart Disease Using Machine Learning Algorithms", International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), pp. 197-202, 2020.
[30] K.M. Al Mustafa, "Prediction of Heart Disease and Classifiers' Sensitivity Analysis", BMC Bioinformatics, Vol. 21, No. 1, pp. 1-18, 2020.

BIOGRAPHIES

Name: **Abbas**
Middle Name: **Fadhil Hamzah**
Surname: **Alharan**
Birthday: 10.03.1985
Birthplace: Babylon, Iraq
Bachelor: Computer Science, Faculty of Science, University of Babylon, Babylon, Iraq, 2007
Master: Computer Science, Computer Science, Faculty of Computing, University Technology Malaysia, Johor Bahru, Malaysia, 2015
The Last Scientific Position: Lecturer, Computer Science Department, Faculty of Education for Girls, University of Kufa, Najaf, Iraq, Since 2019
Research Interests: Machine Learning, Computer Vision, Optimization Problems, Metaheuristic Algorithms

Name: **Luay**
Middle Name: **Mohammed**
Surname: **Saferali**
Birthday: 09.05.1975
Birthplace: Najaf, Iraq
Bachelor: Computer Science, Department of Computer Science, Almansoor University Collage, Baghdad, Iraq, 1997
Master: Information Technology, Faculty of Information Technology, Tenaga University (UniTen), Selangor, Malaysia, 2016
The Last Scientific Position: Lecturer, Computer Science Department, Faculty of Education for Girls, University of Kufa, Najaf, Iraq, since 2016
Research Interests: Computer Security, Network Security, Machine Learning

Name: **Sabah**
Middle Name: **Khudair**
Surname: **Abbas**
Birthday: 06.12.1966
Birthplace: Najaf, Iraq
Bachelor: Department of Computer Science, Faculty of Education for Girls, University of Kufa, Najaf, Iraq, 2006
Master: Computer Science, Faculty of Sciences and Fine Arts, Arts, Science and Technology University, Beirut, Lebanon, 2012
The Last Scientific Position: Assoc. Prof., Imam Al Kadhim University College, Baghdad, Iraq, Since 2022
Research Interests: Artificial Intelligence, Image Processing

Name: **Safaa**
Middle Name: **Jasim**
Surname: **Mosa**
Birthday: 28.10.1981
Birthplace: Baghdad, Iraq
Bachelor: Computer Science, Baghdad College of Economic Sciences University, Baghdad, Iraq, 2003
Master: Computer Science, Information Technology, Iraqi Commission for Computers and Informatics, Informatics Institute for Postgraduate Studies, Baghdad, Iraq, 2006
The Last Scientific Position: Lecturer, Computer Science Department, Faculty of Education for Girls, University of Kufa, Najaf, Iraq, Since 2020
Research Interests: Artificial Intelligence, Image Processing, Database