2024, 21(2 Special Issue): 0812-0831
https://doi.org/10.21123/bsj.2024.9711
P-ISSN: 2078-8665 - E-ISSN: 2411-7986

Baghdad Science Journal

Exploring Important Factors in Predicting Heart Disease Based on Ensemble-Extra Feature Selection Approach

Howida Abubaker [1], ***Farkhana Muchtar*** *[1], ***Alif Ridzuan Khairuddin*** [1], ***Ahmad Najmi Amerhaider Nuar*** [1], ***Zuriahati Mohd Yunos*** [1], ***Carolyn Salimun*** [2]

[1]Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia.
[2]Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia.
*Corresponding Author.

ICAC2023: The 4th International Conference on Applied Computing 2023.

Received 27/09/2023, Revised 10/02/2024, Accepted 12/02/2024, Published 25/02/2024

Abstract

Heart disease is a significant and impactful health condition that ranks as the leading cause of death in many countries. In order to aid physicians in diagnosing cardiovascular diseases, clinical datasets are available for reference. However, with the rise of big data and medical datasets, it has become increasingly challenging for medical practitioners to accurately predict heart disease due to the abundance of unrelated and redundant features that hinder computational complexity and accuracy. As such, this study aims to identify the most discriminative features within high-dimensional datasets while minimizing complexity and improving accuracy through an Extra Tree feature selection based technique. The work study assesses the efficacy of several classification algorithms on four reputable datasets, using both the full features set and the reduced features subset selected through the proposed method. The results show that the feature selection technique achieves outstanding classification accuracy, precision, and recall, with an impressive 97% accuracy when used with the Extra Tree classifier algorithm. The research reveals the promising potential of the feature selection method for improving classifier accuracy by focusing on the most informative features and simultaneously decreasing computational burden.

Keywords: Extra Tree, Feature selection, Feature subsets, Heart Disease Dataset, Machine learning.

Introduction

The Electronic Health Record (EHR) technique has revolutionized the healthcare industry, generating an abundance of clinical data[1-3]. However, this wealth of information poses a significant challenge for disease prediction, as the sheer volume of data and associated features can be overwhelming [4]. This is particularly true for time-sensitive tasks, such as predicting mortality. The healthcare domain is home to numerous medical databases storing vast clinical records, but not all of these are relevant to the predictive task at hand. With the rise of big data and the proliferation of medical records, decision-making based on multiple features has become increasingly complex, with redundant and irrelevant attributes complicating matters [5-8]. These extraneous features introduce noise, leading to inaccuracies in predictions, and increase computational overhead.[9, 10].

Cardiovascular disease datasets are of significant concern, given the high mortality rates associated with heart disease worldwide [8, 9, and 11]. According to the World Health Organization (WHO), the mortality rate of cardiovascular disease is estimated to approach nearly 30 million by the year 2040, as indicated in the studies [12 and 13]. However, diagnosing the disease using extensive datasets is increasingly difficult due to the vast number of features and data samples. Data mining methods offer an efficient means to identify individuals at higher risk of heart disease early on and enhance prediction accuracy. In particular, feature selection methods hold great promise in extracting essential features and discerning patterns from complex healthcare datasets [14, 15]. Additionally, feature selection improves model performance by reducing complexity and increasing prediction accuracy, which is critical in medical diagnosis [16]. Various feature selection methods, such as filter, wrapper, and embedded approaches, have emerged to address the dimensionality challenge[17-19].

An interesting study conducted by [19] employed a variety of feature selection techniques, including principal component analysis, Chi-squared testing ($\chi2$) statistical, relief, and symmetrical, to identify the best feature subsets for predicting heart disease. However, it is worth noting that the experiment was conducted using only one dataset. Furthermore, the chi-square method evaluates features based solely on their association with the class label, which may result in reduced effectiveness when interpreting classifier models. These limitations could potentially cause delays in the learning process and lead to a decline in accuracy performance. Thus, this study utilizes the ensemble-based extra tree feature selection method to identify the most influential features in heart disease prediction. This groundbreaking research holds the potential to significantly improve diagnosis and pave the way for better treatment strategies [20]. The proposed method uses an extra tree classifier's built-in feature importance functionality to determine the relevance of features to the primary classifier. This functionality evaluates the individual significance of each feature in making predictions within the model. This assessment is made by calculating the Gini Index as a measure of feature separation in the data. Features are then ranked based on their relative importance with respect to the class label. Features with higher scores are considered more critical, whereas those with lower scores are less influential in relation to the class target [20] which are more effective and informative features. The significant contributions of this study are outlined as follows:

1. To execute an optimal feature selection model, an ensemble-based extra tree method will be used for optimal feature selection. This method will extract the attributes that possess the highest level of information to enhance the accuracy of predictions and decrease the cost associated with complexity overhead.
2. To build classification models based on a 10-cross-validation approach with the selected features and the full features of the datasets.
3. To investigate the impact of the proposed method on the classification performance by comparing the results of performance metrics of the classification models with selected and full features of heart disease datasets used in this study.

The remainder of the paper is organized as follows: Section 2 provides background information, while Section 3 outlines the research methodology. Section 4 presents the research results, followed by Section 5, which includes discussion. Finally, Section 6 concludes with a summary of findings and future research directions. The focus is on creating an optimal feature selection model that leverages an ensemble of extra trees to extract the most informative attributes. This aims to enhance prediction accuracy while minimizing the complexity overhead cost.

Related Work

Coronary Vascular disease is a global health concern that claims more lives annually than any other disease. In 2016, CVD accounted for nearly 31 percent of deaths, claiming over 17 million lives worldwide. Heart attacks and strokes are the leading causes of this deadly disease [8].

Numerous studies have employed machine learning techniques to diagnose heart diseases with promising results [21]. One such study by done [21] utilized an ensemble of classifiers to predict heart disease using the Cleveland heart dataset from the UCI machine learning repository. Their experiment demonstrated a significant increase in prediction accuracy.

The author in [22] investigated the benefit of using machine learning in predicting heart disease by implementing the improved logistic regression classification model to predict if the patient has heart disease or not.

The study conducted by [23] utilized a variety of machine learning classifier algorithms, including Naïve Bayes, K Nearest Neighbor, Support Vector Machine, Logistic Regression, Extreme Gradient Boost, and Random Forest, to predict heart disease. The Random Forest algorithm produced the highest accuracy rate of 94.12%. However, it should be noted that no feature selection methods were employed to isolate the most important features for diagnosing the disease.

The explosion of medical datasets has yielded high-dimensional data, posing complexities for effective analysis and interpretation, which can cause processing delays and a decrease in the model's classification performance. This is largely due to the presence of irrelevant and redundant features [8]. As a result, many researchers have integrated feature selection methods into the classification process to identify the most impactful features within datasets that influence disease outcomes. By selecting these features, computational overhead can be reduced and accuracy can be enhanced.

For example, in the study of [18], the researchers developed a state-of-the-art predictive system for diagnosing heart disease that combined machine learning and artificial intelligence. This system incorporated seven different classifier algorithms, including logistic regression, K-NN, ANN, SVM, NB, DT, and random forest, along with three feature selection techniques: Relief, mRMR, and LASSO. These techniques were used to identify the most important features for predicting heart disease. The results of their study showed that using the Relief algorithm significantly improved the performance of the logistic regression classifier, achieving an impressive accuracy rate of 89 %. However, it should be noted that the researchers only used one dataset to assess the benefits of feature selection.

In another investigation conducted by [19], they found that the effectiveness of feature selection hinges on the choice of the machine learning classifier. To improve heart disease classification accuracy, researchers explored diverse machine learning and feature selection methods to extract the most valuable features from relevant datasets. Their most precise model attained an accuracy rate of 85.0%, featuring a precision score of 84.73% and a recall rate of 85.56%. This was accomplished by employing Chi-squared feature selection in conjunction with the BayesNet classifier. Additionally, they devised a model utilizing Relief feature selection and the SGD algorithm, which demonstrated comparable accuracy (84.86%) and precision (84.57%), albeit an enhanced recall rate of 85.83%. Another remarkable model combined PCA feature extraction with IBK, resulting in the highest recall rate among all models at 87.22%, accompanied by an accuracy of 83.89% and a precision of 81.91%. It is of utmost importance to underscore the fact that this investigation solely utilized a solitary dataset in order to examine the impact of feature selection on the performance of classification.

In a study investigating heart stroke risk assessment, the author [24] proposed a novel feature selection method named "weighting-and-ranking-based hybrid feature selection" (WRHFS). This technique integrates multiple filter-based approaches, like Information Gain, Fisher score, and standard deviation, to score and rank features. Leveraging prior knowledge, WRHFS identified 9 crucial features out of 28 features for predicting stroke risk. Nevertheless, the approach of information gain (IG) chooses features by considering their significance to the target class, while avoiding the need for classifier models, potentially rendering these features less effective in interpreting the classifier models as highlighted by the study of [25].

The work done by [26] used $\chi 2$ statistical optimum feature selection technique to get the most significant features related to class target in diagnosing heart disease. The result of their experiment shows that accuracy rate increased from 85.29% to 89.7% with the proposed method and SVM classifier algorithm. However, the chi-Square methodology chooses characteristics by considering their connection to the class label, without resorting to classifier models. Consequently, these characteristics do not possess superior effectiveness in deciphering the classifier models [25].

An evaluation of the impact of feature selection on machine learning models for heart disease prediction was conducted by the study of [8]. The ANOVA-F test was employed to select the most critical features from the dataset, with the goal of enhancing prediction accuracy. The experimental results revealed that employing feature selection techniques led to improved performance in machine learning models compared to models utilizing the entire feature set. This not only reduced computational complexities but also enhanced the accuracy of prediction models.

In a related study[7], an ensemble classification model based on a feature selection approach was developed to identify the most relevant features related to the target class. The proposed model achieved an impressive accuracy rate of 97.57% on the datasets they considered. Their findings demonstrated that the utilization of feature selection approaches notably enhanced the performance of the classification model.

The aforementioned findings serve to underscore the efficacy of employing feature selection techniques in augmenting the efficacy of distinct classification algorithms in the context of diagnosing heart disease. They address issues related to noisy features and dependencies within the heart disease dataset that can influence the diagnostic process.

In this paper, we propose the utilization of an ensemble extra tree algorithm feature selection based to select a subset of features for training machine learning algorithms on the datasets to diagnose patients with heart disease. Selecting features with decision tree-based methods is notably quicker and more straightforward when contrasted with techniques like Fisher's score and F-score. A significant drawback of Fisher's score and F-score is their independent feature score calculation, lacking mutual information consideration among features [27]. In contrast, the Extra Trees classifier evaluates all features collectively when categorizing data. This approach acknowledges the potential for certain feature combinations to outperform high-scoring individual features [28, 29], which is why the Extra Trees classifier is used as a feature selector in this study.

Experimental Design

In this study, datasets related to heart disease were obtained from various sources through the Google dataset tool. The acquired datasets underwent two key pre-processing steps: (i) data cleaning, addressing missing and erroneous values; and (ii) data normalization, facilitating optimal performance of machine learning models. The extra tree algorithm was employed for feature selection, extracting the most crucial features relevant to the class target. Subsequently, machine learning models were trained using the selected feature subsets alongside the complete feature dataset.

Evaluation of all classifier models was conducted utilizing diverse evaluation metrics through a 10-fold cross-validation approach, where the dataset was divided into 10 groups. The model was trained using nine of these groups, with the remaining group serving for performance evaluation. This process was iterated 10 times, each time using a different group as the test set during the 10-fold cross-validation. All experiments were implemented using Python and scikit-learn libraries. Fig. 1 provides an illustration of the experiment phases, detailed in subsequent sections.

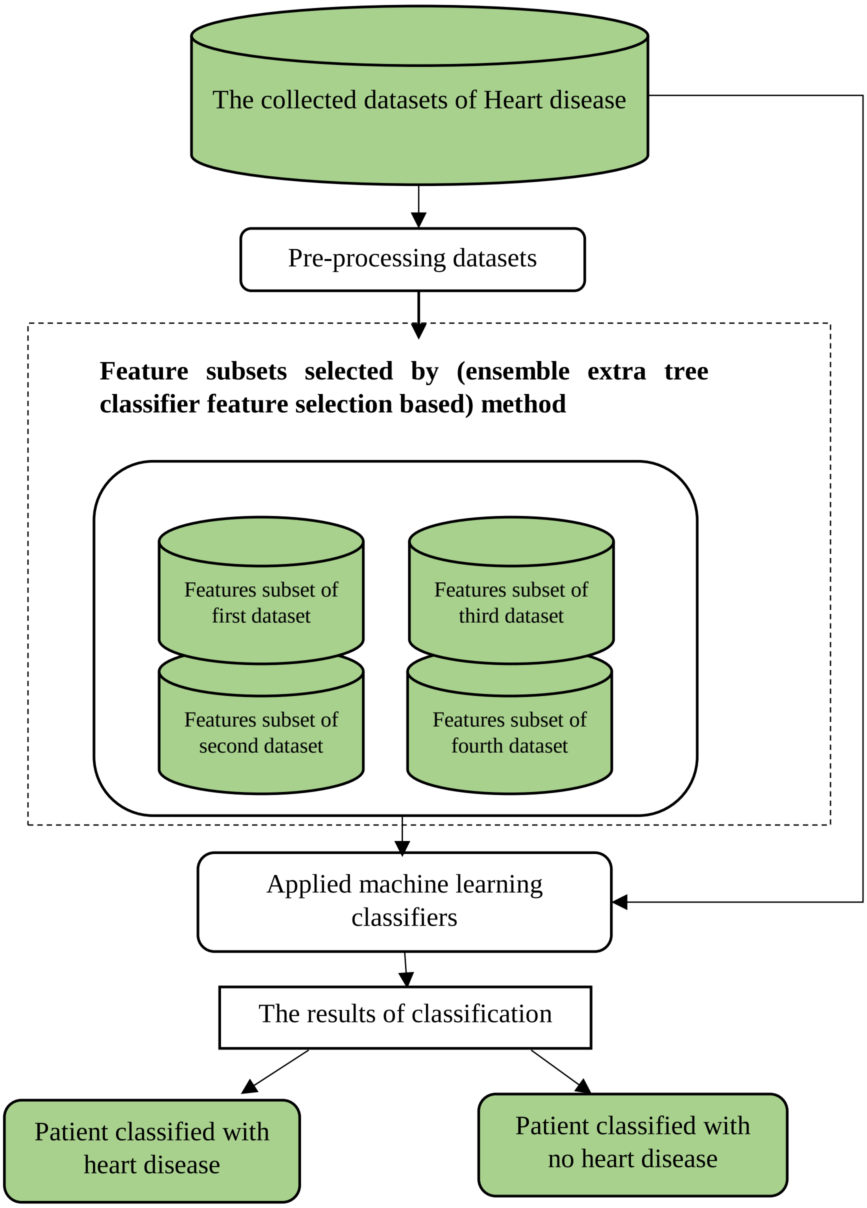

The collected datasets of Heart disease
Pre-processing datasets
Feature subsets selected by (ensemble extra tree classifier feature selection based) method
Features subset of first dataset
Features subset of third dataset
Features subset of second dataset
Features subset of fourth dataset
Applied machine learning classifiers
The results of classification
Patient classified with heart disease
Patient classified with no heart disease

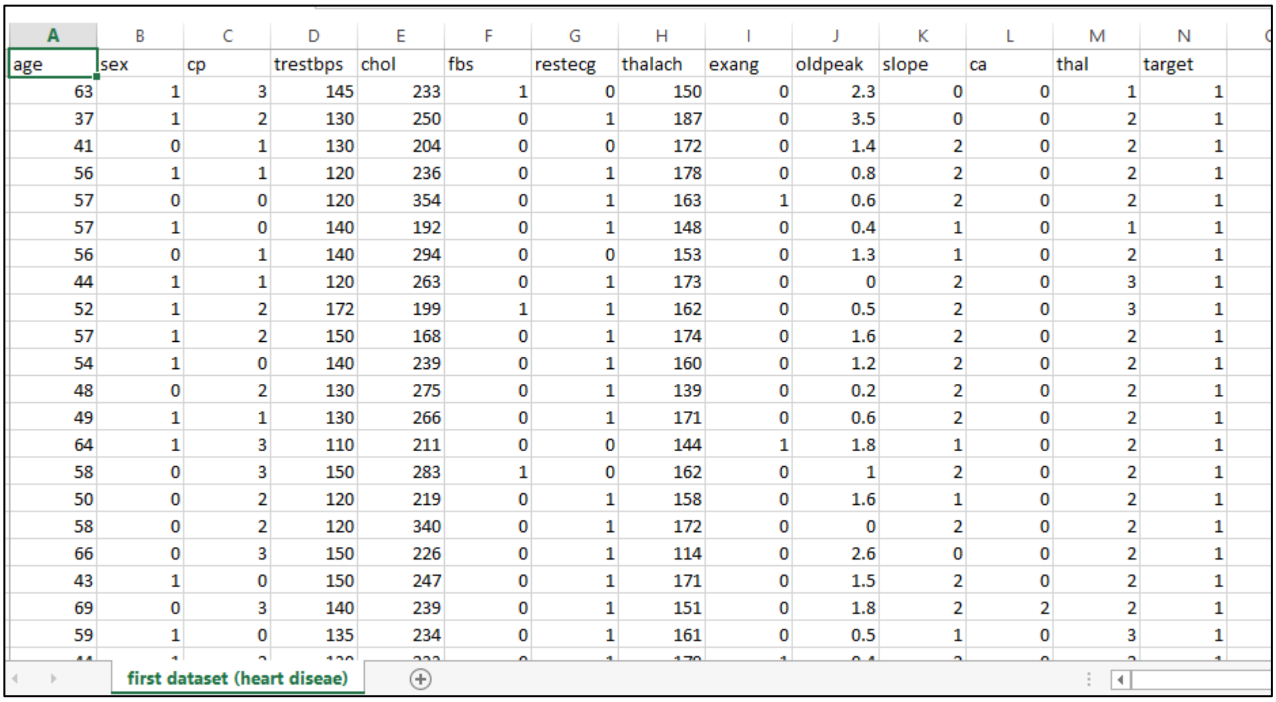
Figure 1. The flow chart of the experiment of the study

Datasets and Features

Four datasets are collected randomly from the Google dataset website. Table 1 presents a description of the datasets used in this study. The first dataset is called Heart-disease and contains 303 samples and 14 attributes (age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, num). The last attribute (num or class label ('target') represents the output of predation in which the presence of heart disease is indicated by 1 and the absence of heart disease is denoted by 0. The dataset was used in the study of [30] and is available on the Kaggle website. The dataset contains 76 attributes, but most previous studies used a subset of 14 features. The explanation of the features for all datasets used in this study is illustrated in detail in Table 2 below [18]. Some samples of this dataset and its attributes are displayed in Fig. 2 below.

Table 1. Description of the datasets used in this study.

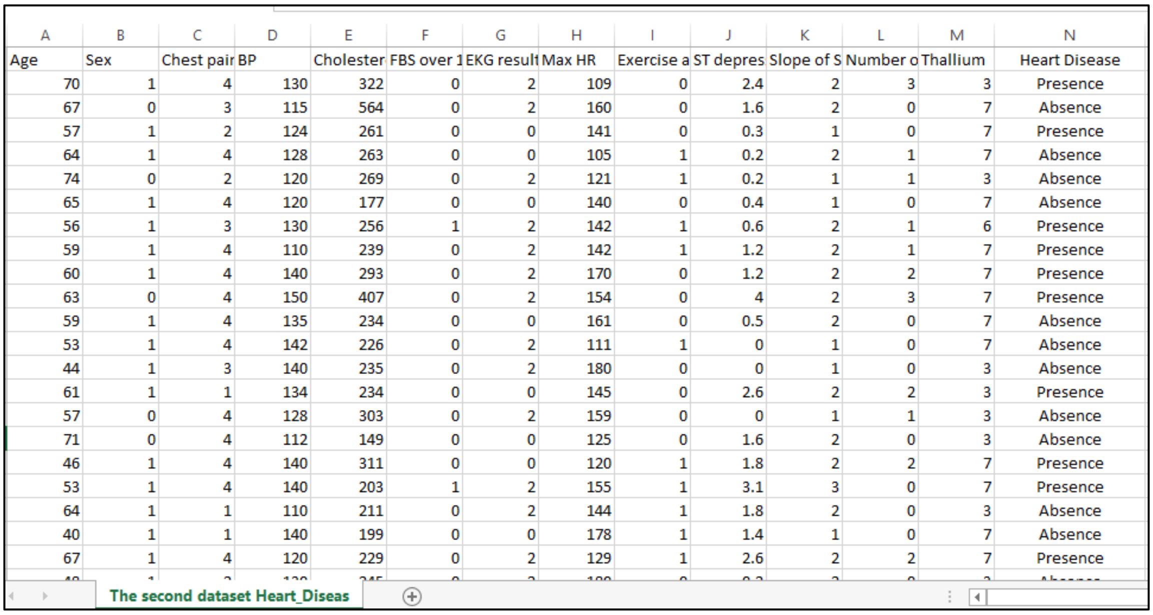
Dataset	Number of features	Number of Samples
Heart-disease	14	303
Heart Disease Prediction	13	270
Heart Disease Dataset	14	1025
Medical dataset	9	1319



age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
58	0	2	120	340	0	1	172	0	0	2	0	2	1
66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
59	1	0	135	234	0	1	161	0	0.5	1	0	3	1

Figure 2. Some samples of Heart Disease dataset

The second dataset is named as **Heart Disease Prediction dataset** which includes 270 patients with 13 independent variables (age, sex, Chest pain type, BP, Cholesterol, FBS over 120, EKG results, Max HR, Exercise angina, ST depression, Slope of ST, Number of vessels fluro, Thallium, Heart Disease) which is available at Kaggle website (Heart Disease Prediction |Kaggle). The “Heart Disease” field refers to the presence of heart disease in the patient; the presence of disease is denoted by “**Presence**” while the absence of disease is indicated by “**Absence**”. Fig. 3 displays some samples of the dataset.

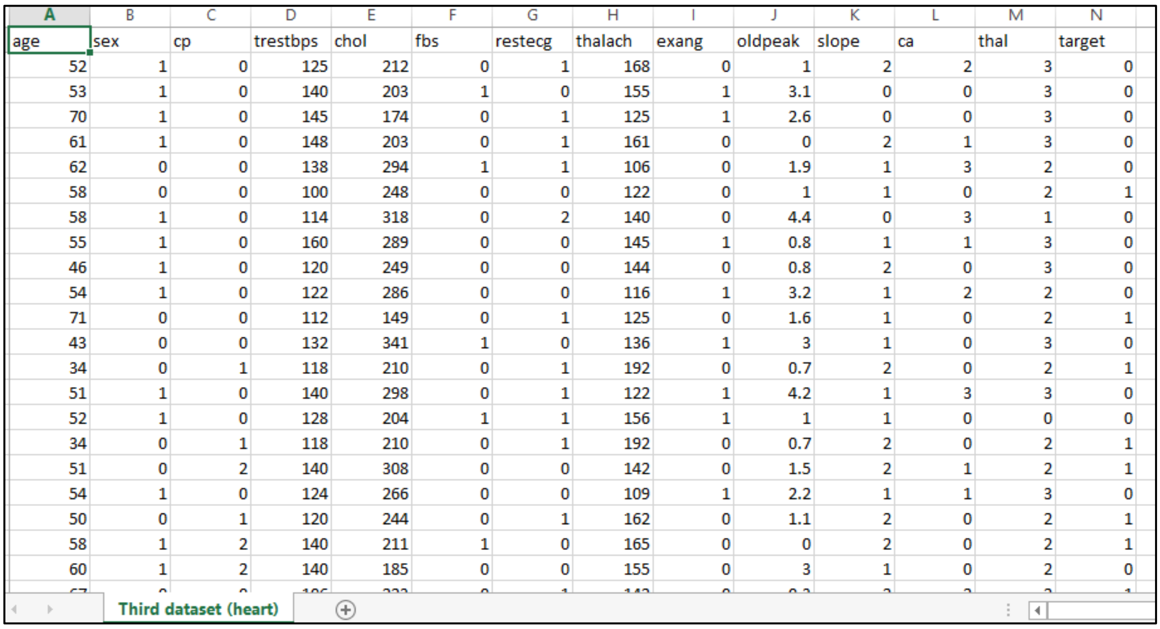


A	B	C	D	E	F	G	H	I	J	K	L	M	N
Age	Sex	Chest pair	BP	Cholester	FBS over 1	EKG result	Max HR	Exercise a	ST depres	Slope of S	Number o	Thallium	Heart Disease
70	1	4	130	322	0	2	109	0	2.4	2	3	3	Presence
67	0	3	115	564	0	2	160	0	1.6	2	0	7	Absence
57	1	2	124	261	0	0	141	0	0.3	1	0	7	Presence
64	1	4	128	263	0	0	105	1	0.2	2	1	7	Absence
74	0	2	120	269	0	2	121	1	0.2	1	1	3	Absence
65	1	4	120	177	0	0	140	0	0.4	1	0	7	Absence
56	1	3	130	256	1	2	142	1	0.6	2	1	6	Presence
59	1	4	110	239	0	2	142	1	1.2	2	1	7	Presence
60	1	4	140	293	0	2	170	0	1.2	2	2	7	Presence
63	0	4	150	407	0	2	154	0	4	2	3	7	Presence
59	1	4	135	234	0	0	161	0	0.5	2	0	7	Absence
53	1	4	142	226	0	2	111	1	0	1	0	7	Absence
44	1	3	140	235	0	2	180	0	0	1	0	3	Absence
61	1	1	134	234	0	0	145	0	2.6	2	2	3	Presence
57	0	4	128	303	0	2	159	0	0	1	1	3	Absence
71	0	4	112	149	0	0	125	0	1.6	2	0	3	Absence
46	1	4	140	311	0	0	120	1	1.8	2	2	7	Presence
53	1	4	140	203	1	2	155	1	3.1	3	0	7	Presence
64	1	1	110	211	0	2	144	1	1.8	2	0	3	Absence
40	1	1	140	199	0	0	178	1	1.4	1	0	7	Absence
67	1	4	120	229	0	2	129	1	2.6	2	2	7	Presence

The second dataset Heart_Diseas

Figure 3. Some samples of Heart Disease Prediction dataset

The third dataset is referred to as the Heart Disease Dataset. It encompasses a total of 1025 observations and 14 attributes. The compilation of this dataset involved the combination of four distinct heart datasets, namely Cleveland, Hungary, Switzerland, and Long Beach V. The dataset incorporates a total of 76 attributes, encompassing the predicted attribute. However, the dataset employs a subset of 14 features. The "target" field refers to the presence of heart disease in the patient and it is indicated by the value (1) while the value (0) refers to the absence of heart disease in the patient. The dataset can be accessed on the Kaggle website (Heart Disease Dataset|Kaggle). Fig. 4 shows some samples of the dataset.

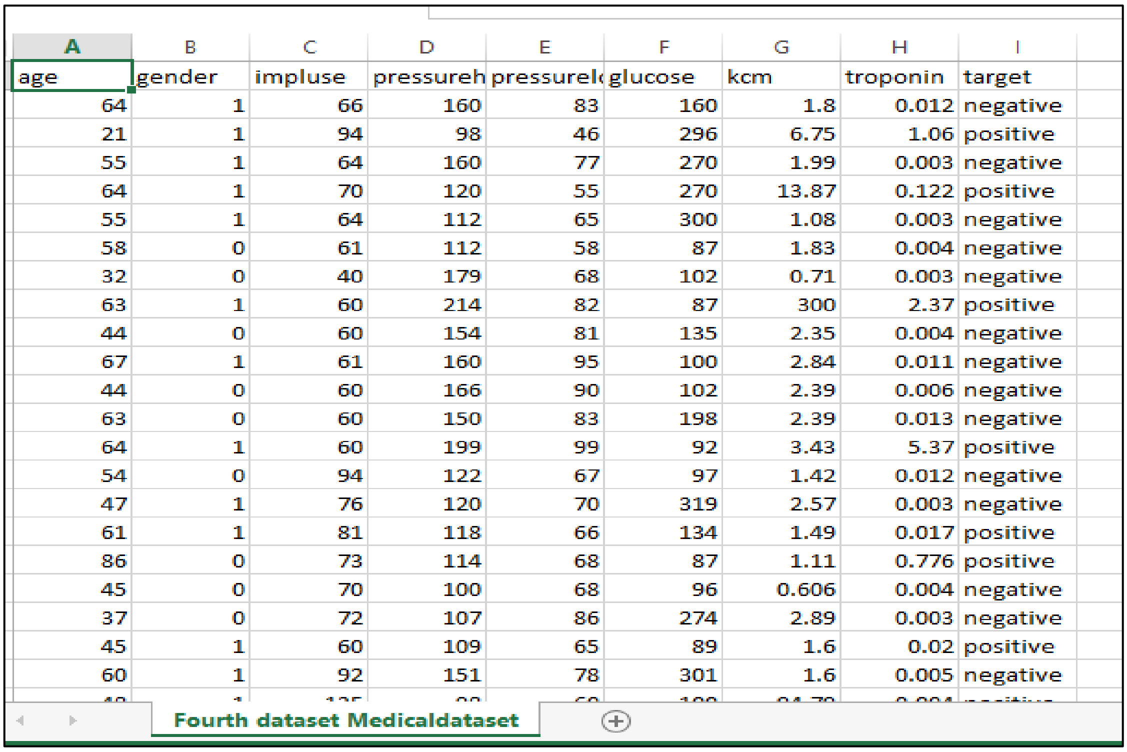


A	B	C	D	E	F	G	H	I	J	K	L	M	N
age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
58	0	0	100	248	0	0	122	0	1	1	0	2	1
58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
43	0	0	132	341	1	0	136	1	3	1	0	3	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
52	1	0	128	204	1	1	156	1	1	1	0	0	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
54	1	0	124	266	0	0	109	1	2.2	1	1	3	0
50	0	1	120	244	0	1	162	0	1.1	2	0	2	1
58	1	2	140	211	1	0	165	0	0	2	0	2	1
60	1	2	140	185	0	0	155	0	3	1	0	2	0

Third dataset (heart)

Figure 4. Some samples of Heart Disease Dataset

The last dataset was the **Medical** dataset used in the study of [22]. This dataset encompassed a total of 1319 samples and contained 9 distinct features. These features included age, gender, heart rate (impulse), systolic blood pressure (pressure high), diastolic blood pressure (pressure low), blood sugar (glucose), CK-MB (kcm), and Test-Troponin (troponin), target. CK-MB (kcm) is an enzyme present in cardiac muscles, and elevated levels of this enzyme in the bloodstream serve as a significant indicator of heart muscle damage. The target attribute represents the output in which negative nominal value signifies the nonexistence of a heart attack, whereas a positive nominal value signifies the manifestation of a heart attack. Fig. 5 depicts some instances of the Medical dataset.



A	B	C	D	E	F	G	H	I
age	gender	impluse	pressureh	pressurel	glucose	kcm	troponin	target
64	1	66	160	83	160	1.8	0.012	negative
21	1	94	98	46	296	6.75	1.06	positive
55	1	64	160	77	270	1.99	0.003	negative
64	1	70	120	55	270	13.87	0.122	positive
55	1	64	112	65	300	1.08	0.003	negative
58	0	61	112	58	87	1.83	0.004	negative
32	0	40	179	68	102	0.71	0.003	negative
63	1	60	214	82	87	300	2.37	positive
44	0	60	154	81	135	2.35	0.004	negative
67	1	61	160	95	100	2.84	0.011	negative
44	0	60	166	90	102	2.39	0.006	negative
63	0	60	150	83	198	2.39	0.013	negative
64	1	60	199	99	92	3.43	5.37	positive
54	0	94	122	67	97	1.42	0.012	negative
47	1	76	120	70	319	2.57	0.003	negative
61	1	81	118	66	134	1.49	0.017	positive
86	0	73	114	68	87	1.11	0.776	positive
45	0	70	100	68	96	0.606	0.004	negative
37	0	72	107	86	274	2.89	0.003	negative
45	1	60	109	65	89	1.6	0.02	positive
60	1	92	151	78	301	1.6	0.005	negative

Figure 5. Some samples of Medical dataset

Table 1. Description of features for datasets used in this study.

Feature	Description	Type
Age	Age in years	Numerical
Sex	Gender	Nominal
cp	Chest pain type	Nominal
trestbps	Resting blood pressure in mmHg	Numerical
chol	Serum cholesterol in mg/dl	Numerical
Fbs	Fasting blood sugar >120 mg/dl Resting	Nominal
restecg	Resting electrocardiographic results	Nominal
thalach	Maximum heart rate achieved	Numerical
exang	Exercise induced angina	Nominal
oldpeak	ST depression induced by exercise relative to rest	Numerical
slope	The slope of the peak exercise ST segment	Nominal

Ca	Number of major vessels colored by fluoroscopy	Nominal
Thal	Heart rate	Nominal
Target, Heart Disease, class	The predicted class: if the patient has heart disease	Numerical, Nominal
Troponin	The level of protein found in the muscles of heart.	Numerical

All the datasets were preprocessed by dealing with missing values and duplicated samples. However, all datasets do not have any missing values or duplicated samples.

The Feature Selection Approach with an Extra Tree Classifier

In this study, an ensemble-based Extra Tree Classifier feature selection method was applied to all datasets used to identify the most discriminative features associated with the target class. Subsequently, new datasets were generated based on the subset of features obtained through the Extra Tree (EX) algorithm approach. Seven different classifier algorithms were employed to build classifier models using these feature subsets. Additionally, classifier models were constructed using both feature subsets and datasets containing all available features, prior to applying the proposed approach. The study involved a comparative analysis between the feature subset dataset selected using the EX method and the dataset containing all features to identify the impact of the proposed method on the classification performance. To evaluate the performance of these classifiers, a 10-fold cross-validation approach was utilized by assigning k with 10 to avoid overfitting issues when evaluating the effectiveness of the classification models. The data is initially separated into 10 parts in which 9 folds are utilized to train the model and the rest one fold is utilized for the testing intention which means that the classifiers were each executed 10 times to ensure that every portion of a split dataset was seen. The specific classification algorithms used will be discussed in detail in the following section.

The Extra Tree Classifier is the abbreviation of extremely randomized trees and is a type of ensemble learning technique that combines the outcomes of multiple decision trees, each of which is constructed from the original training data without subsampling or replacement. It should be noted that this differs from the Random Forest classifier, which employs bootstrap replicas and subsamples the input data with replacement during tree construction. The method uses the "feature importance" of an Extra Tree Classifier to select the relevant features [31]. It calculates impurity-based feature importance, enabling the selection of relevant features while discarding irrelevant ones. Each feature in the dataset is assigned a score between (0 and 1) through feature importance. Higher scores indicate greater relevance to the output variable. This score facilitates the identification of the most important features for model development. Feature importance is a built-in feature in tree-based classifiers and employs a meta-estimator that fits randomized decision trees (also known as extra trees) on various data subsamples. Averaging is used to enhance predictive accuracy and control overfitting while computing feature importance, which is subsequently employed for feature selection. The decision tree carefully examines its options and picks the most impactful feature within the subset. This feature then guides the data split, using the Gini Index to ensure the most informative separation. Averaging is then used to determine important features that contribute to reducing estimate variance and, consequently, are employed in feature selection. Features are ranked in descending order based on their Gini Importance scores, with higher scores signifying greater importance. To select a specific minimum number of top features (e.g., the top 5 features), the "largest (n)" function is employed, where "n" is set to the desired number of features. The extra tree algorithm is used for implementing feature selection in this study for the following reasons:

- Lowering variance: Extra-Tree has a lower variance when comparing to other algorithms due to the averaging of ensemble trees [31].

- Reducing computational complexity and scalability: The algorithm is able to deal with very large-scale data mining applications of a huge number of features and training samples. The algorithm gains speed by using the same procedure repeatedly but sacrifices accuracy by choosing random, non-optimal split points [32 & 33].
- Salient towards non-relevant and redundant inputs: The tree in extra tree algorithm is built to be robust towards non-relevant attributes variables as long as the number of randomized trees is sufficiently large according to the number of samples [32-34].

The Significant Feature Subsets Selected

As previously mentioned, the selection of features was based on analyzing the feature importance property. An extra tree classifier was used to assign scores to input features that were selected using a predictive model that computed the Gini Index. This determined which features were most relevant to the class label, and the ones with the highest scores were deemed the most important while those with the lowest values were considered the least important. The figure below (Fig. 6) highlights the features that hold the most weight in determining the presence or absence of heart disease based on the available dataset (Heart-disease dataset).

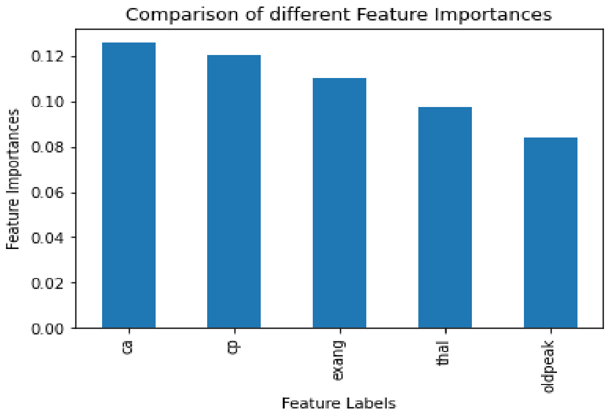


Figure 6. Features subset selected for Heart-disease dataset

As displayed in Fig. 6 above, the number of major vessels colored by the fluoros-copy (ca) attribute represents the most prominent feature that is related to the class object with a score of 0.1258. Chest pain type trestbps attribute comes in the second rank with the weight of 0.1204 followed by exang and Thal features. Oldpeak which represents (ST depression induced by exercise relative to rest) is the least important feature since ranked the last feature. The study done by [9]; declared that the chest pain feature represents the most significant feature in predicting heart disease.

The significant features selected for the second dataset are demonstrated in Fig.7 below.

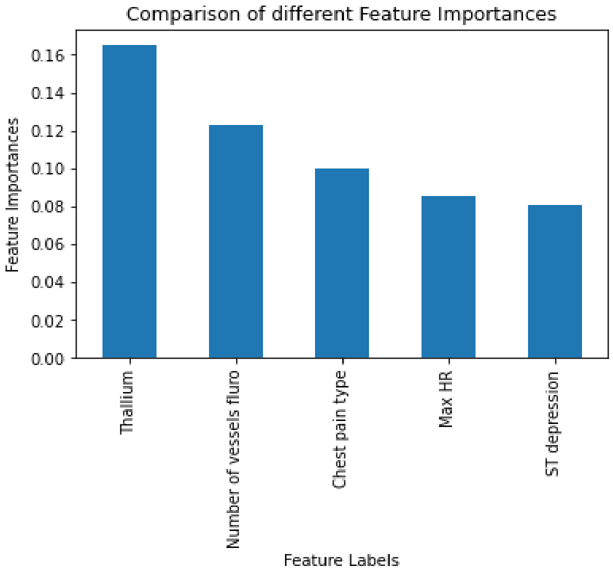


Figure 7. Features subset selected for the second dataset (Heart Disease Prediction)

As illustrated in Fig. 7 above, Thalium has the highest correlation to the possibility of heart disease. As reported by the study of [35], people with Thalium of value 2 are more likely to have heart disease. The number of vessels (CA) is one-factor causing heart disease; people with the lowest value of CA are more likely to have heart disease. Chest pain comes in third rank and as explained by [35 & 36], a serious chest pain condition is considered a significant symptom that causes heart disease. MaxHR which represents the maximum heart rate achieved (Values between 60 and 202); comes in the fourth rank.

According to Fig. 6 and Fig. 7, Oldpeak or ST depression caused by exercise relative to rest feature ranked last in the top of the selected features. Oldpeak attribute ranked in the last list of significant features selected by the study of [9]. Stress or ST depression has a negative effect on a person's heart health which can lead to high blood pressure, arterial damage, irregular heart rhythms, and a weakened immune system [35].

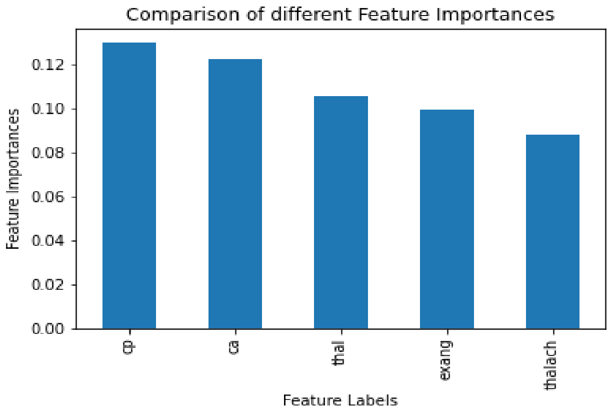


Figure 8. Features subset selected for the third dataset (Heart Disease)

Comparing the selected features of the first and **Heart Disease** datasets as shown in Fig. 6 and Fig. 8 above, some of the optimal selected features are the same since both of the datasets have the same features with different sizes of samples. As observed in Fig. 8 above, chest pain (cp) represents the most important factor in diagnosing patients with heart disease. As reported by [35], chest pain is considered one of the most common symptoms causing a heart attack.

Figure 9 showcases the results of applying the suggested method to the medical dataset. From the initial nine features, the method identified troponin, kcm, age, glucose, and high pressure as the most significant ones. The feature (troponin test) measures the levels of troponin; a type of protein primarily found in the heart muscles, within the bloodstream. Under normal circumstances, troponin is not typically detected in the blood. However, when there is damage to the heart muscles, troponin is released into the bloodstream. The extent of heart damage corresponds to the amount of troponin released. As per the findings of [37], troponin emerged as an independent predictor of cardiovascular-related mortality, myocardial infarction, or stroke in patients afflicted with both type 2 diabetes and stable ischemic heart disease. Additionally, the study done by [38] conducted an experiment demonstrating that troponin was one of the predictor variables associated with heart disease, in alignment with clinical practice and existing literature.

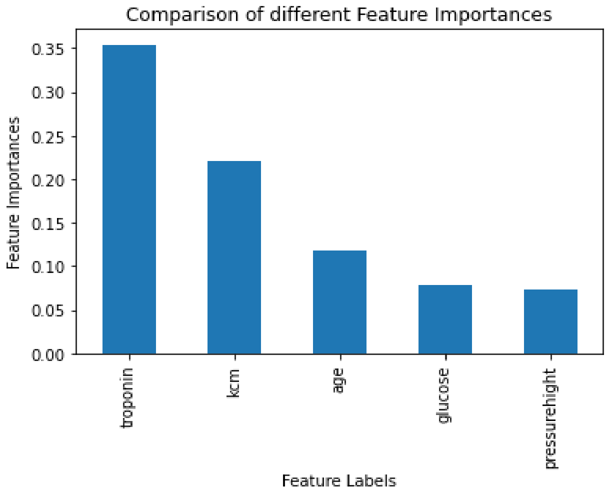


Figure 9. Feature subset selected for fourth dataset (Medical dataset)

Overall, it is concluded from the experiment that chest pain (cp) and the number of vessels colored by fluoroscopy (CA) are the most representative features in predicting heart disease [9.]

Machine Learning Algorithms

To assess the effectiveness of the extracted features obtained through the suggested technique, diverse classifier models have been constructed. These models encompass Support Vector Machine (SVM), k-Neighbors Classifier (KNN), Extra Tree (EX), Naïve Bayes (NB), Linear Discriminant Analysis (LDA), Multilayer Perceptron (MLP), and Logistic Regression (LR). These classifier algorithms have been used in several studies to diagnose patient with heart disease [39-41].The evaluation process employed a 10-fold cross-validation, where each classifier was executed 10 times to ensure comprehensive coverage of all sections within the split dataset, as outlined by [42]. The performance of these classifiers is then compared against two benchmarks: the best features selected using the proposed method and the full features set. To implement the classification models, the following parameters are used:

For SVM:

- Kernel: The kernel serves to specify the type of kernel that is to be employed in the algorithm. A variety of kernel types exist, including, but not limited to, "linear," "poly," "rbf," "sigmoid," and "precomputed." For the purposes of this study, the linear kernel has been selected.

For KNN:

- N_neighbors: n_neighbors represent the number of neighbors to be used by default for k neighbors queries, in this experiment, parameter value 7 is used.

For NB:

- There are different types of NB, in this study, Gaussian NB is used with default parameters.

For MLP:

- Hidden_layer_sizes:Hidden_layer_sizes represents the number of neurons in the ith hidden layer, in this study, value 10 is used for Hidden_layer_sizes.

For the other classifier algorithms, the default parameters are used.

Evaluation Metric

To evaluate performance of the ML models, the following evaluation criteria are used:

- Accuracy: The metric reflecting the success rate of a classification model, measured as the ratio of correct predictions (true positives and true negatives) to the total number of predictions.

$$\text{Accuracy} = (TP+TN / TP+TN+FP+FN) \qquad 1$$

The true positive (TP) in this study represents the patient with heart disease correctly classified. And the false positive (FP) represents the patient with heart disease wrongly identified as the patient without heart disease. The true negative (TN) indicates the patient who has no heart disease correctly classified. And the false negative (FN) represents the patient who has no heart disease wrongly classified as a patient with heart disease.

- Precision: In the context of classification, precision refers to the proportion of correctly identified positive instances relative to all instances that the classifier has identified as positive. It is an important metric that evaluates the accuracy of a classifier's positive predictions. A high precision score indicates that the classifier has a low rate of false positives, which is desirable in many business and academic settings. Precision is a key performance indicator for classifiers and plays a vital role in evaluating their efficacy.

$$\text{Precision} = (TP / TP+FP) \qquad 2$$

- Recall: In the context of classification, recall is defined as the ratio of true positive instances to the total number of actual positive instances. In other words, it refers to the proportion of positive cases that are correctly identified by the classifier out of all the instances that are actually positive. This metric is an important measure of the effectiveness of a classifier in identifying relevant instances, particularly in domains where false negatives can have serious consequences. Therefore, it is crucial to evaluate the recall rate of a classifier along with other performance metrics, such as precision and F1 score, to obtain a comprehensive understanding of its performance.

$$\text{Recall} = (TP / TP+FN) \qquad 3$$

All the performance metrics are averaged since the experiment was carried out using 10 fold cross-validation method.

Results and Discussion

The Classification Results

This section provides a detailed explanation of the experiment outcomes. The performance results of the machine learning algorithms tested on the final feature subsets obtained through the extra tree feature selection proposed approach, based on all features used in this study, are presented in the Table. Additionally, the Table displays the performance results of the tested ML models on the datasets with full feature subsets, to evaluate the effectiveness of the proposed feature selection method on the classification performance. The classification performance results for the heart disease dataset can be found in Table 3 below.

According to the data in Table 3, the MLP classifier performed the best when utilizing our proposed feature selection method, achieving an impressive **85%** precision, recall, and accuracy rate for each metric. This represents a significant improvement compared to not using the feature selection technique, where all three metrics were at 74%. Additionally, the KNN classifier also showed enhanced results with our feature selection approach, reaching **83%** precision, recall, and accuracy for each metric compared to the full feature results of 67%. These findings demonstrate the effectiveness of feature selection algorithms in accurately classifying heart disease with fewer relevant features, ultimately improving the overall classification performance. Nonetheless, the other classifiers demonstrated comparable or near-comparable classification performance utilizing the chosen features in comparison to the complete features set of the datasets employed. These results confirm that utilizing a reduced number of crucial features is preferable to utilizing all features in order to minimize complexity overhead.

Table 3. The performance result of classification models for Heart-disease dataset on the selected and all features.

Classifier algorithms	Number of Features	Precision	Recall	Accuracy
SVM	Selected features	0.83	0.83	0.83
	All features	0.84	0.83	0.83
KNN	Selected features	**0.83**	**0.83**	**0.83**
	All features	0.67	0.67	0.67
NB	Selected features	0.81	0.81	0.81
	All features	0.81	0.81	0.81
EX	Selected features	0.81	0.81	0.81
	All features	0.83	0.83	0.83
LDA	Selected features	0.83	0.82	0.82
	All features	0.83	0.82	0.82
MLP	Selected features	**0.85**	**0.85**	**0.85**
	All features	0.74	0.74	0.74
LR	Selected features	0.84	0.83	0.83
	All features	0.83	0.83	0.83

Table 4 presents a summary of the performance metrics for the **Heart Disease Prediction** dataset. The most significant features chosen for this dataset are Thallium, Number of vessels fluoroscopy, Chest pain type, Max HR, and ST depression. The proposed feature selection method has resulted in a feature subset that exhibits excellent classification precision at **84%**, along with a recall and accuracy of **84%** when using the LR classifier. When utilizing the selected features with the KNN classifier, the classification performance has improved significantly from **67%** to **75%**. Additionally, the MLP classifier has achieved more favourable results with the reduced features set, achieving a rate of **83%** compared to a rate of 77% with the full set of features. These results unequivocally demonstrate that machine learning models outperform models that use the entire feature set, underscoring how this feature selection approach enhances prediction model accuracy while reducing the feature space's dimension. However, the other classifiers showed similar or nearly identical classification performance when using the selected features in comparison to the full feature set of the datasets. These outcomes validate the preference for using a smaller set of essential features to reduce complexity and computational overhead.

Table 4 .The performance result of classification models for Heart Disease Prediction dataset on the selected and full features.

Classifier algorithms	Number of Features	Precision	Recall	Accuracy
SVM	Selected features	0.83	0.83	0.83
	All features	0.83	0.83	0.83
KNN	Selected features	**0.75**	**0.75**	**0.75**
	All features	0.67	0.67	0.67
NB	Selected features	0.84	0.84	0.84
	All features	0.84	0.84	0.84
EX	Selected features	0.80	0.80	0.80
	All features	0.83	0.83	0.83
LDA	Selected features	0.84	0.84	0.84
	All features	0.84	0.84	0.84
MLP	Selected features	**0.83**	**0.83**	**0.83**
	All features	0.77	0.77	0.77
LR	Selected features	**0.84**	**0.84**	**0.84**
	All features	0.83	0.83	0.83

Table 5 showcases the results of the experiment on the Heart Disease Dataset. The KNN classifier displayed superior performance in precision, recall, and accuracy, achieving a remarkable **81%** for each metric when utilizing the reduced set of features (including chest pain (cp), the number of vessels colored by fluoroscopy (CA), Thal, thalach, and exang) as recommended by our method. In contrast, when using the full set of features, the KNN classifier exhibited the lowest performance metrics, with precision, recall, and accuracy all at **72%**. These results suggest that the proposed feature selection approach effectively enhances disease prediction accuracy. However, it is observed that the NB classifier produced performance metrics of 79% when using the selected features, compared with 82% when using all features. Additionally, the EX and MLP classifiers produced lower metric values, with **97%** and **81%** respectively, when using the reduced feature set, compared to **100%** and **83%** respectively when using all features. These results indicate that using fewer discriminative features achieves similar or close classification performance compared to utilizing all features. In conclusion, reducing the feature size by using fewer features is a better option than using all features to minimize complexity overhead costs [43-45].

Table 5. The performance result of classification models for Heart Disease Dataset on the selected and all features.

Classifier algorithms	Number of Features	Precision	Recall	Accuracy
	Selected features	0.81	0.81	0.81
SVM	All features	0.85	0.84	0.84
	Selected features	**0.81**	**0.81**	**0.81**
KNN	All features	0.72	0.72	0.72
	Selected features	0.79	0.79	0.79
NB	All features	0.82	0.82	0.82

EX	Selected features	0.97	0.97	0.97
	All features	1.00	1.00	1.00
LDA	Selected features	0.81	0.80	0.80
	All features	0.84	0.83	0.83
MLP	Selected features	0.81	0.81	0.81
	All features	0.83	0.83	0.83
LR	Selected features	0.81	0.81	0.81
	All features	0.84	0.84	0.84

Table 6 below displays the performance metrics obtained from the experiment conducted on the **Medical** dataset for both selected features and full features. It was observed that classifiers EX and LR produced better results when using reduced features as compared to their corresponding classifiers which utilized all features. For instance, the EX classifier recorded the highest precision, recall, and accuracy scores of **97%, 97%**, and **97%** respectively, when using selected features as compared to other classifiers. The LR classifier also achieved better precision, recall, and accuracy scores of **81%, 80 %**, and **80 %** respectively, when using selected features as compared to full features. Furthermore, utilizing feature selection approaches improved the classification performance for each metric with reduced features using the MLP classifier, increasing from **73%** to **75%**. The results indicated that using a fewer number of significant features is useful in classifying objects, resulting in improved results and reduced computational cost overhead [5, 43, 45]. In conclusion, some classifiers show better results when using all features, but the difference is negligible as compared to using selected features [1, 18, 45].

Table 6. The performance result of classification models for the Medical dataset on the selected and all features.

Classifier algorithms	Number of Features	Precision	Recall	Accuracy
SVM	Selected features	0.82	0.81	0.81
	All features	0.82	0.81	0.81
KNN	Selected features	0.63	0.63	0.63
	All features	0.63	0.64	0.64
NB	Selected features	0.82	0.68	0.68
	All features	0.82	0.68	0.68
EX	Selected features	**0.97**	**0.97**	**0.97**
	All features	0.92	0.92	0.92
LDA	Selected features	0.71	0.71	0.71
	All features	0.70	0.70	0.70
MLP	Selected features	**0.74**	**0.75**	**0.75**
	All features	0.73	0.73	0.73
LR	Selected features	**0.81**	**0.80**	**0.80**
	All features	0.79	0.79	0.79

Study Comparison with Earlier Works

This study conducted a comparative analysis of various feature selection methods employed in predicting heart disease. Table 7 displays a comprehensive comparison of the outcomes obtained from this study and previous ones. The results demonstrate that the suggested ensemble extra tree feature selection method outperformed other techniques, exhibiting an impressive accuracy rate of 97%.

Table 7. Proposed model comparative analysis with current state-of-the-art methods.

Study	Feature Selection	Classifier	Total features	Acc (%)	Selected Features	Acc (%)
[18]	Relief, mRMR, and LASSO	logistic regression, K-NN, ANN, SVM, NB, DT, and random forest	13	-	6	89
[26]	Chi-squared (χ2)	SVM	14	85.29	6	89.7
[39]	Information Gain	KNN, NB and RF	13	-	10	95.63
[44]	Weighting- and ranking-based hybrid feature selection (WRHFS)	-	28	-	9	81.5
Proposed method	An ensemble extra tree classifier feature selection based	SVM, KNN, EX, NB, LDA, MLP, and LR	9	92	5	97

Discussion

After analysing the tables presented above, it is clear that selecting fewer important features can improve the classification performance. For example, the Heart-disease and Heart-disease prediction dataset's accuracy rates increased from **0.67** to **0.83** and **0.76**, respectively.

Table 3 demonstrates that the MLP classifier achieved the highest accuracy rate of **0.85** with the reduced features for the Heart-disease dataset. Moreover, among the classifier algorithms tested on the Medical dataset, the extra tree classifier, when trained with the optimal reduced feature set, exhibited the highest accuracy **(97%)**, surpassing alternative approaches.

In conclusion, the experimental findings suggest that utilizing only relevant features with specific classifier algorithms can significantly enhance machine learning model performance [46-48]. This approach yields comparable or nearly identical results to models that utilize the full feature set across all datasets. The study recommends utilizing a reduced number of features, and the results emphasize the impact of feature selection techniques in decreasing the feature space's size, improving machine learning model performance, and lowering complexity overhead cost. Analysis indicates that even with a limited number of features, machine learning models display superior performance compared to models that use the full features set.

Conclusion and Future Work

The development of massive clinical data has led to a significant challenge for disease prediction due to the huge volume of data and associated features. These features could be redundant and irrelevant, and do not provide significant information to the predictive task and can decay the accuracy of the disease prediction. Therefore, the principal aim of this paper is to investigate the impact of the feature selection method in improving the classification performance by decreasing the feature space size by proposing an extra tree feature selection-based approach to identify the most crucial features that are close to the class target and enhance the prediction accuracy. This investigation was carried out by utilizing set of features obtained from commonly utilized heart disease datasets accessible via the Google dataset tool. Experiments were undertaken to investigate the impact of the suggested approach for feature selection on the performance of prediction. The experiments were carried out in two scenarios, one with feature selection and the other without feature selection. Four datasets, varying in sample sizes and features, were employed for these experiments. The analysis encompassed seven classifier algorithms: SVM, KNN, EX, NB, LDA, MLP, and LR classifiers. Furthermore, the models underwent evaluation based on precision, recall, and accuracy metrics. The Extra Tree model outperformed all other models when using the selected features, achieving an accuracy rate of 97 % on the Medical dataset. The experimental outcomes underscore the promise of employing an extra tree feature selection-based technique to extract the most distinctive features that help in diagnosing patients with heart disease or not with similar or close classification performance results compared to using all features with evident enhancements in classification performance. However, some classifiers with the reduced features performed similar or close results compared with the full features. So, it is recommended in the future to use some statistical tests to find out a suitable classifier that achieves better results in predicting heart disease with the most optimal selected features. In addition, other feature selection approaches used for extracting the most important features in predicting heart disease could be explored in future work.

Authors' Declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been included with the necessary permission for re-publication, which is attached to the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee in Universiti Teknologi Malaysia.

Authors' Contribution Statement

H.A proposed the presented idea, performed the computations process, verified the analytical methods, proofed the outcome of the experiments, and contributed to the writing of the manuscript. All authors discussed the results and contributed to the final manuscript.

References

1. Joloudari JH, Saadatfar H, Dehzangi A, Shamshirband S. Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection. Inform Med Unlocked. 2019; 17(August):100255. https://doi.org/10.1016/j.imu.2019.100255
2. Sarwade JM, Mathur H. Performance analysis of symptoms classification of disease using machine learning algorithms. 2020; 11(3):2024–2032. https://doi.org/10.1109/ICECA55336.2022.10009407
3. Elzeheiry HA, Barakat S, Rezk A. Different Scales of Medical Data Classification Based on Machine Learning Techniques: A Comparative Study. Applied Sciences (Switzerland). 2022; 12(2). https://doi.org/10.3390/app12020919
4. Alelyani S. Stable bagging feature selection on medical data. J Big Data. 2021; 8(1). https://doi.org/10.1186/s40537-020-00385-8
5. Liu S, Yao J, Zhou C, Motani MS. SURI: Feature Selection Based on Unique Relevant Information for Health Data. Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018. 2019; 687–692. https://doi.org/10.1109/BIBM.2018.8621163
6. Zhang F, Luo C, Lan C, Zhan J. Benchmarking feature selection methods with different prediction models on large-scale healthcare event data. TBench. 2021; 1(1):100004. https://doi.org/10.1016/j.tbench.2021.100004
7. Abdollahi J, Nouri-Moghaddam B. A hybrid method for heart disease diagnosis utilizing feature selection based ensemble classifier model generation. Sci Iran D Comput Sci Eng Electr Eng. 2022; 5(3):229–246. https://doi.org/10.1007/s42044-022-00104-x
8. Pathan MS, Nag A, Pathan MM, Dev S. Analyzing the impact of feature selection on the accuracy of heart disease prediction. Healthc Anal (N Y). 2022; 2(April):100060. https://doi.org/10.1016/j.health.2022.100060
9. Joloudari JH, Joloudari EH, Saadatfar H, Ghasemigol M, Razavi SM, Mosavi A, Nadai L. Coronary artery disease diagnosis; ranking the significant features using a random trees model. Int J Environ Res Public

Health. 2020; 17(3). https://doi.org/10.3390/ijerph17030731

10. Pavithra V, Jayalakshmi V. Review of feature selection techniques for predicting diseases. Proceedings of the 5th International Conference on Communication and Electronics Systems, ICCES 2020. 2020; 1213–1217. https://doi.org/10.1109/ICCES48766.2020.9138058
11. Dissanayake K, Johar MGM. Comparative study on heart disease prediction using feature selection techniques on classification algorithms. Applied Computational Intelligence and Soft Computing. 2021. https://doi.org/10.1109/ICREST51555.2021.9331158
12. Bashir S, Khan ZS, Hassan Khan F, Anjum A, Bashir K. Improving Heart Disease Prediction Using Feature Selection Approaches. Proceedings of 2019 16th International Bhurban Conference on Applied Sciences and Technology, IBCAST 2019. 2019; 619–623. https://doi.org/10.1109/IBCAST.2019.8667106
13. Roberts CA, Binder M, Antoine D. Reflections on Cardiovascular Disease: The Heart of the Matter. In: Binder M, Roberts CA, Antoine D, editors. The Bioarchaeology of Cardiovascular Disease. Cambridge: Cambridge University Press; 2023. p. 258–62. https://www.cambridge.org/core/books/abs/bioarchaeology-of-cardiovascular-disease/reflections-on-cardiovascular-disease/23D4812D144A7823B844EEA28559E5F4
14. Latha CBC, Jeeva SC. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Inform Med Unlocked. 2019; 16(June):100203. https://doi.org/10.1016/j.imu.2019.100203
15. Das S, Sultana M, Bhattacharya S, Sengupta D, De D. XAI–reduct: accuracy preservation despite dimensionality reduction for heart disease classification using explainable AI. J Supercomput. 2023. https://doi.org/10.1007/s11227-023-05356-3
16. Azmi J, Arif M, Nafis MT, Alam MA, Tanweer S, Wang G. A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. Med Eng Phys. 2022; 105(May):103825. https://doi.org/10.1016/j.medengphy.2022.103825
17. Saw M, Saxena T, Kaithwas S, Yadav R, Lal N. Estimation of prediction for getting heart disease using logistic regression model of machine learning. 2020 International Conference on Computer Communication and Informatics, ICCCI 2020. 2020; 20–25. https://doi.org/10.1109/ICCCI48352.2020.9104210
18. Haq AU, Li JP, Memon MH, Nazir S, Sun R, Garciá-Magarinõ IA. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems. 2018. https://doi.org/10.1155/2018/3860146
19. Spencer R, Thabtah F, Abdelhamid N, Thompson M. Exploring feature selection and classification methods for predicting heart disease. Digit Health. 2020; 6:1–10. https://doi.org/10.1177/2055207620914777
20. Kharwar AR, Thakor D V. An Ensemble Approach for Feature Selection and Classification in Intrusion Detection Using Extra-Tree Algorithm. Int J Inf Secur Priv. 2022; 16(1):1–21. https://www.igi-global.com/gateway/article/285019
21. Firdaus FF, Nugroho HA, Soesanti I. A Review of Feature Selection and Classification Approaches for Heart Disease Prediction. IJITEE (International Journal of Information Technology and Electrical Engineering). 2021; 4(3):75. https://doi.org/10.22146/ijitee.59193
22. Maghdid S, Rashid TA. An Extensive Dataset for the Heart Disease Classification System. Mendeley Data. 2022. https://doi.org/ 10.17632/65gxgy2nmg.2
23. Bora N, Gutta S, Hadaegh A. Using Machine Learning to Predict Heart Disease. WSEAS Trans. Biol. Biomed. 2022; 19:1–9.https://www.semanticscholar.org/paper/Using-Machine-Learning-to-Predict-Heart-Disease-Bora-Gutta/3f531cbc6abe322151382d69b25f1a4559867a44
24. Ahmad GN, Fatima H, Abbas M, Rahman O. Mixed Machine Learning Approach for Efficient Prediction of Human Heart Disease by Identifying the Numerical and Categorical Features. 2022. https://www.mdpi.com/2076-3417/12/15/7449
25. Cai L, Li Y, Xiong Z. JOWMDroid: Android malware detection based on feature weighting with joint optimization of weight-mapping and classifier parameters. Comput Secur. 2021 Jan 1; 100:102086. https://doi.org/10.1016/j.cose.2020.102086
26. Sarra RR, Dinar AM, Mohammed MA, Abdulkareem KH. Enhanced Heart Disease Prediction Based on Machine Learning and χ2 Statistical Optimal Feature Selection Model. Designs. 2022; 6(5). https://doi.org/10.3390/designs6050087
27. Hamidzadeh J. Robust Feature Selection by Filled Function and Fisher Score. 2022. https://doi.org/10.21203/rs.3.rs-1102788/v1
28. Alfian G, Syafrudin M, Fahrurrozi I, Fitriyani NL, Atmaji FTD, Widodo T, et al. Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method. Comput.

2022; 11(9). https://doi.org/10.3390/computers11090136

29. Sanmorino A, Marnisah L, Sunardi H. Feature Selection Using Extra Trees Classifier for Research Productivity Framework in Indonesia. In: Lecture Notes in Electrical Engineering. 2023. p. 13–21. https://doi.org/10.1007/978-981-99-0248-4_2
30. Yazdani A, Varathan KD, Chiam YK, Malik AW, Wan Ahmad WA. A novel approach for heart disease prediction using strength scores with significant predictors. BMC Med Inform Decis Mak. 2021; 21(1):1–16. https://doi.org/10.1186/s12911-021-01527-5
31. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning. 2006; 63(1):3–42. https://doi.org/10.1007/s10994-006-6226-1
32. Moosmann F, Triggs B, Jurie F. Fast discriminative visual codebooks using Randomized Clustering Forests. Adv Neural Inf Process Syst. 2007; 985–92. https://ieeexplore.ieee.org/document/6287461
33. Sharma J, Giri C, Granmo OC, Goodwin M. Multi-layer intrusion detection system with ExtraTrees feature selection, extreme learning machine ensemble, and softmax aggregation. Eurasip J. Inf. Secur. 2019 Dec; 2019(1):1-6.https://doi.org/10.1186/s13635-019-0098-y
34. Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding variable importances in forests of randomized trees. Advances in neural information processing systems. 2013; 26. https://dl.acm.org/doi/10.5555/2999611.2999660
35. Scirica BM, Bhatt DL, Braunwald E, et al. Prognostic Implications of Biomarker Assessments in Patients with Type 2 Diabetes at High Cardiovascular Risk: A Secondary Analysis of a Randomized Clinical Trial. JAMA Cardiol. 2016; 1(9):989–998. https://doi.org/10.1001/jamacardio.2016.3030
36. Bradley J, Schelbert EB, Bonnett LJ, Lewis GA, Lagan J, Orsborne C, et al. Predicting hospitalisation for heart failure and death in patients with, or at risk of, heart failure before first hospitalisation: a retrospective model development and external Validation study. Lancet Digit Health. 2022; 4(6):e445–e454. https://doi.org/10.1016/S2589-7500(22)00045-0
37. Abawajy J, Darem A, Alhashmi AA. Feature subset selection for malware detection in smart IoT platforms. Sensors (Basel). 2021; 21(4):1–19. https://doi.org/10.3390/s21041374
38. Sulaiman Maghdid S. Analysis and Prediction of Heart Attacks Based on Design of Intelligent Systems. J Mech Contin Math Sci. 2019; 14(4). https://doi.org/10.26782/jmcms.2019.08.00051
39. Rajesh N, Maneesha T, Hafeez S, Krishna H. Prediction of heart disease using machine learning algorithms. Int J Eng Technol. 2018; 7(2.32 Special Issue 32):363–6. https://doi.org/10.14419/ijet.v7i2.32.15714
40. Enad HG, Mohammed MA. A Review on Artificial Intelligence and Quantum Machine Learning for Heart Disease Diagnosis: Current Techniques, Challenges and Issues, Recent Developments, and Future Directions. Fusion Pract Appl. 2023; 11(1):08–25. https://doi.org/10.54216/FPA.110101
41. Ayon SI, Islam MM, Hossain MR. Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques. IETE J Res. 2022; 68(4):2488–507. https://doi.org/03772063.2020.1713916
42. Christopher JJ, Nehemiah HK, Arputharaj K, Moses GL. Computer-assisted Medical Decision-making System for Diagnosis of Urticaria. MDM Policy Pract. 2016; 1(1). https://doi.org/10.1177/2381468316677752
43. Elgin Christo VR, Khanna Nehemiah H, Minu B, Kannan A. Correlation-based ensemble feature selection using bioinspired algorithms and classification using backpropagation neural network. Comput Math Methods Med. 2019. https://doi.org/10.1155/2019/7398307
44. Zhang Y, Zhou Y, Zhang D, Song W. A stroke risk detection: Improving hybrid feature selection method. J. Med. Internet Res. 2019; 21(4). https://doi.org/10.2196/12437
45. Ali F, El-Sappagh S, Islam SMR, Kwak D, Ali A, Imran M, et al. Ali F, El-Sappagh S, Islam SR, Kwak D, Ali A, Imran M, Kwak KS. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Inf Fusion. 2020 Nov 1; 63:208-22. https://doi.org/10.1016/j.inffus.2020.06.008
46. Bharti R, Khamparia A, Shabaz M, Dhiman G, Pande S, Singh P. Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. Comput Intell Neurosci. 2021; 2021. https://doi.org/10.1155/2021/8387680
47. Mohammad AM, Attia H, Ali YH. Comparative Analysis of MFO, GWO and GSO for Classification of Covid-19 Chest X-Ray Images. Baghdad Sci J. 2023; 20(January 2020):1540–58. https://doi.org/10.21123/bsj.2023.9236
48. Mahmood RAR, Abdi AH, Hussin M. Performance evaluation of intrusion detection system using selected features and machine learning classifiers. Baghdad Sci J. 2021; 18(2):884–98. https://doi.org/10.21123/bsj.2021.18.2(Suppl.).0884

2024, 21(2 Special Issue): 0812-0831
https://doi.org/10.21123/bsj.2024.9711
P-ISSN: 2078-8665 - E-ISSN: 2411-7986

Baghdad Science Journal

استكشاف العوامل المهمة في التنبؤ بأمراض القلب بناءً على نهج اختيار الميزات الإضافية

هويدا أبو بكر[1]، فرحانة مختار[1]، أليف رضوان خير الدين[1]، أحمد نجمي أميرحيدر نوار[1]، زريعاتي محمد يونس [1]، وكارولين سالمون [2]

[1]كلية الحاسبات، الجامعة التكنولوجية الماليزية، جوهور، ماليزيا.

[2]كلية الحوسبة والمعلوماتية، جامعة ماليزيا صباح، جالان UMS، 88400 كوتا كينابالو، صباح، ماليزيا.

الخلاصة

تعد أمراض القلب حالة صحية خطيرة ومؤثرة، وتصنف على أنها السبب الرئيسي للوفاة في العديد من البلدان. من أجل مساعدة الأطباء في تشخيص أمراض القلب والأوعية الدموية، تتوفر مجموعات البيانات السريرية كمرجع. وبالرغم منذلك، مع ظهور البيانات الضخمة ومجموعات البيانات الطبية، أصبح من الصعب بشكل متزايد على الممارسين الطبيين التنبؤ بدقة بأمراض القلب بسبب وفرة الميزات غير ذات الصلة والمتكررة التي تعيق التعقيد والدقة الحسابية. على هذا النحو، تهدف هذه الدراسة إلى تحديد الميزات الأكثر تمييزًا ضمن مجموعات البيانات عالية الأبعاد مع تقليل التعقيد وتحسين الدقة من خلال تقنية تعتمد على اختيار ميزات الشجرة الإضافية. تقوم الدراسة بتقييم أداء خوارزميات التصنيف المختلفة على أربع مجموعات بيانات حسنة السمعة، باستخدام كل من مجموعة الميزات الكاملة والمجموعة الفرعية للميزات المخفضة المحددة من خلال الطريقة المقترحة. أظهرت النتائج أن تقنية اختيار الميزات تحقق دقة تصنيف ودقة واسترجاع متميزة، بدقة مذهلة تبلغ 97% عند استخدامها مع خوارزمية مصنف Extra Tree. تسلط هذه النتائج الضوء على إمكانات نهج اختيار الميزات في تعزيز أداء التصنيف من خلال تحديد الميزات الأكثر أهمية وذات صلة بالفئة المستهدفة مع تقليل عبء التعقيد الحسابي.

الكلمات المفتاحية: شجرة إضافية، اختيار الميزات، مجموعات فرعية من الميزات، مجموعة بيانات أمراض القلب، التعلم الآلي.