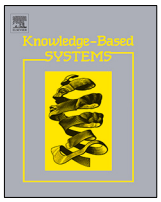
Knowledge-Based Systems 255 (2022) 109709

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



ETCD: An effective machine learning based technique for cardiac disease prediction with optimal feature subset selection

Savita Wadhawan [a,b,*], Raman Maini [b]

[a] MMICTBM, MM(DU), Mullana, Ambala, India
[b] Department of CSE, Punjabi University, Patiala, India

ARTICLE INFO

Article history:
Received 4 February 2022
Received in revised form 2 August 2022
Accepted 13 August 2022
Available online 28 August 2022

Keywords:
UCI repository
Machine learning
Optimal Feature Subset Selection (OFSSA)
ETCD
Support Vector Machine (SVM)
k Nearest Neighbour (kNN)

ABSTRACT

Cardiac disease is the leading cause of death worldwide. The early diagnosis and prognosis can help patients live longer by lowering mortality and boosting survival rates. The paucity of radiologists and doctors in various nations, due to a variety of factors, is a substantial barrier to early diagnosis. Computational intelligence is an emerging concept in the field of medical imaging to identify, prognosticate, and diagnose disease, among numerous initiatives to construct decision support systems. It relieves radiologists and doctors from being overworked and reduces the time it takes to diagnose patients promptly. In this work, an effective technique for cardiac disease (ETCD) prediction based on machine intelligence has been proposed. To ensure the success of our proposed model, we used effective Data Collection, Data Pre-processing, and feature selection process to generate accurate data for the training model. ETCD utilizes the optimal feature subset selection algorithm (OFSSA) to extract features from different datasets (Cleveland, Hungarian, Combined dataset, and Z_Alizadeh Saini datasets) having varying properties available at the UCI machine learning repository. With ETCD, the average accuracy performance for considered datasets gets increased with SVM, KNN, DT, NB, and RF classifiers by 6.227%, 2.72%, 7.345%, 14.084%, and 18.921% respectively. Further, the results of the experiments demonstrate that ETCD outperformed several contemporary baseline approaches in terms of accuracy and was comparable in terms of sensitivity, specificity, precision, and F_Score. ETCD returns the best feasible solution among all input predictive models considering performance criteria and improves the efficacy of the system, hence can assist doctors and radiologists in a better way to diagnose cardiac patients.

1. Introduction

According to the World Health Organization (WHO), cardiac disease (CD) is one of the world's deadliest diseases, responsible for the majority of deaths [1]. CD is caused by a condition in which the heart fails to pump enough blood to other parts of the body, resulting in heart failure [2]. Coronary artery blockage is the most common cause of heart failure. Irregular heartbeat, shortness of breath, chest pain or discomfort, swelling feet or ankles, exhaustion, and fainting are early symptoms of CD. A patient's life expectancy can be extended through early diagnosis and prognosis. A key bottleneck in this regard is the lack of resources and unavailability of doctors in developing or low-income nations, which leads to disease diagnosis at an advanced stage. This is one of the main reasons why almost half of the cardiac patients only live for 1–2 years [3]. The patient's medical history, age, sex, and lifestyle are all risk factors for Cardiac Disease. By changing the lifestyle, such as increased physical activity and avoiding smoking, can help to minimize risk factors by lowering cholesterol and blood pressure. Early detection, changes in lifestyle, and medical examination reports from medical specialists all aid in the diagnosis of the disease.

Although the patient records are usually examined by experts, highly dependent upon expert knowledge. Hence, the probability of human mistakes makes precision and prognostication impossible [4]. Therefore, the requirement of a high level of expertise is one of the prime reasons for the researcher's inclination toward an automated solution that can help in simplifying the diagnosis process. For early detection of cardiac disease, Various ML-based expert systems have been developed [5–10] by researchers. The procedures to implement these ML-based systems include data collection, data pre-processing, model selection, parameter tuning, model training, and testing, model evaluation, and prediction. These machine learning methods, which identify hidden associations from clinical records, are used to detect or even forecast disease progression. Machine learning algorithms can be beneficial in diagnosing diseases when trained on adequate data. Public

* Corresponding author at: MMICTBM, MM(DU), Mullana, Ambala, India.
E-mail addresses: savitawadhawan@mmumullana.org (S. Wadhawan), ramanmaini@pbi.ac.in (R. Maini).

https://doi.org/10.1016/j.knosys.2022.109709
0950-7051/

datasets on cardiac disease are available for comparing prediction models. The development of machine learning and artificial intelligence aids researchers in developing the best prediction model possible using the enormous databases available. Because the clinical datasets may consist of inconsistent and redundant records, appropriate pre-processing is essential [11]. Also, choosing a database's most important attributes might improve the performance of machine learning algorithms [12]. Several studies have been published in the literature that used various feature selection techniques and datasets to detect cardiac disease [9,12–16]. After getting the relevant features, appropriate classifiers and hybrid models can be applied to detect the diseases. To develop classifiers and hybrid models, researchers used a variety of techniques [5,13,17]. Still, several issues, such as feature subset selection, machine learning algorithm implementations, a dearth of in-depth analysis, and the use of limited medical datasets, may obstruct accurate cardiac disease prediction. This study addresses some of the research gaps to develop a better model for the prediction of cardiac disease. In this research, we work with different datasets like the Cleveland dataset, Hungarian dataset, combined dataset (in which four datasets from the UCI repository were combined to get a large dataset), and Z_Alizadeh Saini dataset. Four Feature ranking methods ReliefF, Infogain, Chi-Square, Correlation-based feature selection along with proposed OFSSA were utilized to select the most appropriate features which help to deal with overfitting and underfitting problems of machine learning. Further, SVM (Support Vector Machine), KNN (K-Nearest Neighbours), DT (Decision Tree), NB (Naïve Bayes), and RF (Random Forest) classifiers were utilized. Working with a variety of datasets, Feature selection methods, and classifiers move towards the generalization of the model. Therefore, the study introduces an Effective machine learning based technique for Cardiac Disease Prediction (ETCD) to improve both performance accuracy as well as the most accurate prediction.

Main Contribution: The following are some of the important research contributions:

(i) Proposed a new feature subset selection algorithm OFSSA (Optimal feature subset selection algorithm) utilizing the attribute rank from different feature ranking algorithms.
(ii) Proposed an effective machine learning based technique for cardiac disease prediction (ETCD), returns the best features using OFSSA, and improves the efficacy of various classifiers.
(iii) Comparison of the performance of original dataset vs balanced dataset.
(iv) Compare the performance of "top N features" of various feature ranking methods RelieF, info gain, Chi-square, and Correlation-based FS with proposed OFSSA for different classifiers.
(v) Performance comparison of the proposed ETCD with respect to other state of art methods.
(vi) Compare and contrast the outcome of various datasets of diverse nature (Cleveland (303 records with 14 attributes), Hungarian (294 records with 14 attributes), Combined dataset (920 records with 14 attributes), and Z_Alizadeh Saini dataset (303 records with 56 attributes) w.r.t performance metric (Acc, Sens, Spec, Precision, F_Score).

The rest of the paper is organized as follows: The associated work is included in Section 2 along with a gap analysis. Section 3 provides a detailed description of the materials and methods used. In Section 4, the suggested ETCD framework is presented. The experimental and performance outcomes are shown in Section 5, and the conclusion is presented in Section 6.

2. Related work

In recent years, machine learning algorithms have grown in popularity as a way to boost prediction accuracy and efficiency [18]. The most crucial feature of research in this field is the capacity to produce and choose models with the maximum degree of efficiency and accuracy [18]. Hybrid models, which integrate many machine learning models with essential components, are one viable approach for disease prediction [19].

According to the research conducted by [17] on 92 papers published between 2010–2020, and [20] on 149 publications published for the prognosis of Cardiovascular Diseases (CVDs) between 2000 and 2015, it was found that machine learning and data mining-based algorithms are widely used for feature selection and classification. Recently published studies have employed a variety of publicly accessible datasets. Machine learning and traditional techniques like random forest (RF), support vector machine (SVM), and learning models were recently explored on the UCI Heart Disease dataset [21]. The voting-based strategy enhanced accuracy when used in conjunction with multiple classifiers. According to the study, an improvement of 2.1 percent was achieved for anaemic classifiers [22]. The author in [7] proposed a model using ICA with meta-heuristic for feature selection and KNN for classification with an accuracy of 94.43%± 6.25%. Similarly [10] developed a machine intelligence framework MIFH that used FAMD for feature selection with various classifiers and found FAMD+RF to be a good classifier with 93.44% accuracy. The author in [23] produced a novel heterogeneous hybrid feature selection (2HFS) algorithm, employed SMOTE and ADASYN to deal with data imbalance, and used DT, GNB, RF, and XGBoost as a classifier to achieve maximum accuracy of 92.58% for the Z-Alizadeh dataset followed by 83.94% and 81.58% accuracy for Hungarian and Va datasets respectively. A two-tier ensemble-based coronary heart disease (CHD) detection model in which RF, GB (gradient boosting machine), and extreme gradient boosting machine as ensemble learners were proposed by Tama et al. [6]. The suggested model outperformed in CHD detection in terms of accuracy, F1, and AUC, with values of 98.13%, 96.6%, and 98.7%, respectively. Further in [8], the author suggested the HDPM prediction model, which used DBSCAN for outlier identification, SMOTE_ENN for data balancing, XGBoost for classification, and achieved 98.40% and 95.9% accuracy for Cleveland and Statlog data respectively. In another study, Different machine learning classification algorithms were employed to predict chronic disease, [24]. In their investigation, the Hoeffding classifier predicted CVD with an accuracy of 88.56 percent. For prediction [24] employed both the individual and ensemble learning algorithms techniques such as KNN, J48, Bayes Net, Random Tree, Naïve Bayes, multilayer perceptron, and random forest. J48 was the most accurate with a score of 70.77%. Then, they used cutting-edge approaches with KERAS achieving an accuracy rate of 80%. The ensemble technique was used in [25] to increase prediction accuracy. The accuracy of weak classifiers was improved using bagging and boosting techniques, and the performance for risk identification of heart disease was rated good. They used the majority vote of Bayes Net, Multilayer Perceptron, C 4.5, Nave Bayes, Random Forest (RF), and PART classifiers to create the hybrid model. The designed model attained an accuracy of 85.48 [15] discovered the critical risk factors, applied machine learning models (NB, KNN, LR, DT, SVM, Neural Network, and a hybrid of voting with NB and LR), and provided a comparative analysis. Their research [15] revealed that the hybrid model, when combined with the selected attributes, attained an accuracy of 87.41%. Mohan et al. [5] used a variety of feature combinations as well as various well-known classification approaches. The suggested HRFLM achieved an accuracy of 88.7% using an ANN

with backpropagation and 13 clinical attributes as input. Along with this, SVM, NN, KNN, and DT algorithms were examined, and proven that SVM is beneficial to improve disease prediction accuracy. Dinesh et al. [26] analysed 920 records from the UCI machine learning repository (Cleveland, Hungarian, Switzerland, and Long-Beach-Va) and produced 80.89% accuracy with Random forest. Vijayashree and Sultana [14] suggested a heart disease classification approach that combines the PSO and SVM, with a classification accuracy of 84.36%. Purushottam et al. [27] developed a rule-based classifier for heart disease prediction that was 86.7% accurate. Whereas, the author of [28] suggests using a machine-learning-based prediction and classification system to identify future values of linked vital signs for both chronic respiratory and cardiovascular disorders. Gradient Boosting Models have been repeatedly shown to be one of the most effective techniques for creating predictive models [29,30], but they may cause overfitting, overemphasize outliers, and be computationally expensive because they frequently require many trees, which can be memory and time-intensive. **Gap Analysis** According to the literature, machine learning models were successfully used for predicting cardiac illness, and the majority of investigations were conducted using arbitrary datasets that are available on the UCI machine learning repository. Similarly, various ranking methods are available for feature selection and it is challenging to determine which method is appropriate for a considered dataset. To the best of our knowledge, no study was presented which works well for cardiac disease prediction with the best feature selection for datasets having varying properties for different classifiers. Therefore, designing an effective machine learning based framework for CD diagnosis is a key contribution of this study. The research contribution in the proposed work includes designing a machine learning based framework for cardiac disease prediction which will be independent of feature selection methods and the classifier used. To test the efficacy of the proposed framework, publicly available cardiac disease datasets (Cleveland, Hungarian, Combined dataset, and Z_Alizadeh Saini datasets) available on the University of California Irvine (UCI) repository have been used. Since all datasets consist of mixed type features (nominal, numeric, and binary), pre-processing has been done to make the dataset complete and ready for processing. Then, OFSSA is utilized to choose pertinent features. The suggested framework uses the machine learning classifiers SVM, kNN, DT, NB, and RF to classify normal and heart patients. The performance metric (Acc, Sens, Spec, Precision, F_Score) is used to evaluate the framework performance.

3. Material and methods

The materials and methods used in our experiment are described in this section. It includes information on datasets used, data pre-processing followed by feature selection and a conceptual framework for detecting cardiac disease.

3.1. Datasets

3.1.1. Cleveland/Hungarian dataset

Cleveland and Hungarian datasets are available on the UCI (University of California, Irvine) Repository. Originally, these datasets consist of a total of seventy-six features out of which only fourteen features including class labels have been selected for experimentation because many of them repeat similar information, and some attributes are not related to the target attribute. The considered features are AG (Age), RBS (Resting Blood Sugar), SCH (Serum Cholesterol), MHR (Max. Heart Rate achieved), and STDER (ST depression induced by exercise relative to rest) are numeric in nature. SX(Sex), FBS (Fasting Blood Sugar), and EIG (Exercise-induced angina) are binary features. CPT (Chest pain type), RELR (Resting electrocardiographic result), SPES (Slope of the peak exercise ST segment), MVCF (Number of major vessels coloured by fluoroscopy), and DT (Thallium Defect Type) are nominal in nature. The feature HD is taken as the target feature has 5 levels depending on angiographic disease status such as 0-Healthy, 1-diagnosed with stage 1, 2-diagnosed with stage 2, 3-diagnosed with stage 3, 4-diagnosed with stage 4. In this study, we consider whether or not a person has been diagnosed with cardiac disease. Therefore, label 0 is considered a normal patient, and labels 1–4 are considered a cardiac patient. A detailed description of all considered features is given in Table 1. Cleveland dataset consists of 303 records, out of which 164 are normal patients and 139 are cardiac patients. Similarly, the Hungarian dataset consists of 294 records, out of which 188 are normal patients and 106 are cardiac patients.

3.1.2. Combined dataset

This dataset is formed by combining four different datasets (Cleveland, Hungarian, Switzerland, and Long-Beach-Va datasets) available on the UCI repository. A total of 920 patients record are present in this dataset, out of which, 411 belong to class 0, which is normal and 509 are having Cardiac Diseases. The purpose of combining these datasets into one is to get a large dataset for a more accurate outcome. The features description of this dataset is given in Table 1.

3.1.3. Z-Alizadeh Sani dataset

Z-Alizadeh Sani dataset used in this study consists of information on 303 patients with 56 features, out of which, 216 patients have CD, and 87 patients are normal patients. This dataset contains four different types of features that are demographics (AG, WT, LN, SX, BMI, DM, HTN, CuS, ExS, FH, OB, CRF, CVA, AirD, ThD, CHF, DPL), symptoms, and examination (BP, PR, ED, WPP, LR, SyMu, DiMu, TCP, Dys, FnC, AtP, NoanCP, ExCP, LoTHA), electrocardiogram (ECG) (Rhy, QW, STEL, STDP, TIN, LVH, PRWP), laboratory and echo features (FBS, Cr, TG, LDL, HDL, BUN, ESR, HB, K, Na, WBC, Lymph, Neut, PLT, EF, RRWMA, VHD) described in Table 2. The feature CATH categorizes the CD from Normal. The diameter narrowing above 50% represents a patient as CD, and its absence is stated as Normal [12].

3.2. Data pre-processing

3.2.1. Handling missing values

Collecting entire information from the subject is difficult or almost impossible in the real-life scenario because of disruptions in data flow, privacy issues, or the patient's unwillingness to cooperate. Therefore, the medical datasets consisting missing information on the features as well. There were three types of features found in the original datasets: numeric, binary, and nominal as presented in Tables 1 and 2. The statistical descriptions emphasize the presence of missing values in the original datasets. In a feature F_i, the significance level for missing values is set to 45% [31]. Experimentally it was observed that if a considerable amount of an attribute's data is missing, such as more than 45 percent, it may influence performance or give impartial findings and there is little to no difference in the outcome. The attributes were filled according to algorithm 1, to make the dataset complete and ready for processing. Binary attributes are filled with constant value, nominal attributes are filled with majority value of that attribute and the numeric attributes are filled with mean of that attribute.

Table 1
Detail description of Cleveland, Hungarian and Combined datasets.

Feature no.	Feature name	Feature description	Feature type	Domain (Cleveland dataset)	Domain (Hungarian dataset)	Domain (Combined dataset)
1	AG	Age in Years	Numeric	[29 77]	[28 66]	[29 77]
2	SX	Sex {1=male, 0=female}	Binary	[0 1]	[0 1]	[0 1]
3	CPT	Chest pain type {1: typical angina;2: atypical angina, 3: non-anginal pain, 4: asymptomatic}	Nominal	[1 4]	[1 4]	[1 4]
4	RBS	Resting Blood Sugar (mm Hg)	Numeric	[94 200]	[92 200]	[0 200]
5	SCH	serum cholesterol in mg/dl (mg/dl)	Numeric	[126 564]	[85 603]	[85 603]
6	FBS	fasting blood sugar > 120 mg/dl {1 = true; 0 = false}	Binary	[0 1]	[0 1]	[0 1]
7	RELR	Resting electrocardiographic result {0: normal, 1: ST-T wave abnormality, 2: showing probable or definite left ventricular hypertrophy}	Nominal	[0 2]	[0 2]	[0 2]
8	MHR	Max. Heart rate achieved	Numeric	[71 202]	[82 190]	[60 202]
9	EIG	Exercise induced angina {1 = yes; 0 = no}	Binary	[0 1]	[0 1]	[0 1]
10	STDER	ST depression induced by exercise relative to rest	Numeric	[0 6.2]	[0 5]	[-0.5 6.2]
11	SPES	Slope of the peak exercise ST segment {1: upsloping, 2: flat, 3: downsloping}	Nominal	[1 3]	[1 3]	[1 3]
12	MVCF	Number of major vessels coloured by fluoroscopy {0-3}	Nominal	[0 3]	[0 0]	[0 3]
13	DT	Thallium Defect Type {3 = normal;6 = fixed defect; 7 = reversible defect}	Nominal	[3 7]	[3 7]	[3 7]
14	HD	Diagnosis of heart disease (angiographic disease status) {Value 0: <50% diameter narrowing(Normal), value 1: >50% diameter narrowing(Patient)}	Binary	[0 4]	[0 4]	[0 4]

Algorithm 1: HMV // Handling missing values

```
Input:
    D: {F1, F2, ..., Fn}   //D: Dataset with missing values featurs Fi (Incomplete features).
Output:
    D= {F1, F2, ..., Fm}   where m<=n   //D: Dataset consisting each feature Fi as a complete feature.
begin
    for ∀ Fi
        If #missing values(Fi) > 45% of total values
            Remove Fi from dataset.
        else
            If Fi _type = binary
                Put a constant value in w.r.t all missing values.
            endif
            If Fi _type = nominal
                Put majority value of Fi w.r.t all missing values.
            endif
            If Fi _type = numeric
                Put mean of Fi w.r.t all missing values.
            endif
        endif
    endfor
end
```

3.2.2. Data normalization

To eliminate numerical inconsistencies during the computational process, data normalization was performed after handling missing values. One of the most prominent data normalization methods, min–max normalization, was picked out of various normalization methods. By using Eq. (1), the value λ was mapped to λ' in the range $[n_{min}, n_{max}]$ in the min–max normalization method.

$$\lambda' = n_{min} + [n_{max} - n_{min}] * \frac{\lambda - \lambda_{min}}{\lambda_{max-}\lambda_{min}} \quad (1)$$

Where, $[n_{min}, n_{max}]$ is the attribute range. The data was declared smooth and ready for FS after handling missing values and normalization.

3.2.3. Data balancing

Machine Learning methods utilized the term "Imbalanced Data Distribution" to characterize the situation in which observations in one class are either significantly greater or lower than the those in other classes. Standard machine learning algorithms tend to just consider the majority class and ignore the minority, which significantly misclassifies the minority class. Due to their imbalance nature, cardiac disease datasets need to be balanced [8]. Here, we use SMOTE (synthetic minority oversampling technique) to balance the cardiac disease dataset, which is one of the most popular oversampling methods that balance class distribution by recreating minority class cases at random. It creates new minority instances by combining existing minorities. It creates virtual training records for the minority class using linear interpolation. These synthetic training records are created for each example in the minority class by selecting one or more of the k-nearest neighbours at random according to pseudocode 1. The data is subsequently reconstituted, after which classification models can be applied to the processed data.

4. Proposed methodology

4.1. Feature selection

For machine learning algorithms, feature selection techniques play a very important role to choose the best features and these selected features help to reduce the execution time. Further, the selection of the best features has a significant impact on medical data analysis to get a quick and accurate diagnosis. For feature selection, we have two kinds of approaches: first, using filter methods that select features based on their relationship with their target, and second, by using wrapper methods that utilize the learning algorithm itself to estimate the values of the features. Various type of filter methods is available in literature like Info gain, chi-square test, fisher score, relief, Correlation-based feature selection (CFS), and so on. All these algorithms have different characteristics and working principles. Some work well with binary datasets and others are good for multiclass datasets. Along with this, the different feature gets different rank according to the working principle of the technique as shown in Table 6. Table 6 shows the rank given to all attributes of the above-mentioned datasets with considered ranking methods. It is observed that no feature receives the same rank as all techniques. Therefore, it is challenging to determine which strategy is appro-

Table 2
Detail description of the Z-Alizadeh Sani dataset.

Sr. no.	Feature name	Feature description	Feature type	Range
1	AG	Age	Numeric	[30 86]
2	WT	Weight	Numeric	[48 120]
3	LN	Length	Numeric	[140 188]
4	SX	Sex {1=male, 0=female}	Binary	[0 1]
5	BMI	Body Mass Index (Kb/m2)	Numeric	[18 41]
6	DM	Diabetes mellitus	Binary	[0 1]
7	HTN	Hypertension	Binary	[0 1]
8	CuS	Current smoker	Binary	[0 1]
9	ExS	Ex-smoker	Binary	[0 1]
10	FH	Family history	Binary	[0 1]
11	OB	Obesity {Yes if MBI > 25, No otherwise}	Binary	[0 1]
12	CRF	Chronic Renal Failure	Binary	[0 1]
13	CVA	Cerebrovascular Accident	Binary	[0 1]
14	AirD	Airway disease	Binary	[0 1]
15	ThD	Thyroid disease	Binary	[0 1]
16	CHF	Congestive heart failure	Binary	[0 1]
17	DPL	Dyslipidemia	Binary	[0 1]
18	BP	Blood pressure (mm Hg)	Numeric	[90 190]
19	PR	Pulse Rate ppm	Numeric	[50 110]
20	ED	Edem	Binary	[0 1]
21	WPP	Weak peripheral pulse {Yes, No}	Binary	[0 1]
22	LR	Lung rates{Yes, No}	Binary	[0 1]
23	SyMu	Systolic murmur {Yes, No}	Binary	[0 1]
24	DiMu	Diastolic murmur {Yes, No}	Binary	[0 1]
25	TCP	Typical chest pain {Yes, No}	Binary	[0 1]
26	Dys	Dyspnea {Yes, No}	Binary	[0 1]
27	FnC	Function class {1, 2, 3, 4}	Nominal	[0 3]
28	AtP	Atypical {Yes, No}	Binary	[0 1]
29	NoanCP	Nonanginal chest pain {yes, No}	Binary	[0 1]
30	ExCP	Exertional chest pain {yes, No}	Binary	[0 1]
31	LoTHA	Low TH Ang(low-threshold angina) {yes, No}	Binary	[0 1]
32	Rhy	Rhythm {Sin, AF}	Binary	[0 1]
33	QW	Q wave	Binary	[0 1]
34	STEL	ST elevation	Binary	[0 1]
35	STDP	ST depression	Binary	[0 1]
36	TIN	T inversion	Binary	[0 1]
37	LVH	LVH (left ventricular hypertrophy) {Yes, No}	Binary	[0 1]
38	PRWP	Poor R-wave progression {Yes, No}	Binary	[0 1]
39	FBS	Fasting blood sugar (mg/dL)	Numeric	[62 400]
40	Cr	Creatine (mg/dL)	Numeric	[0.5 2.2]
41	TG	Triglyceride(mg/dL)	Numeric	[37 1050]
42	LDL	Low-density lipoprotein (mg/dL)	Numeric	[18 232]
43	HDL	High-density lipoprotein (mg/dL)	Numeric	[15 111]
44	BUN	Blood urea nitrogen (mg/dL)	Numeric	[6 52]
45	ESR	Erythrocyte sedimentation rate (mm/h)	Numeric	[1 90]
46	HB	Hemoglobin (g/dL)	Numeric	[8.9 17.6]
47	K	Potassium (mEq/lit)	Numeric	[3.0 6.6]
48	Na	Sodium (mEq/lit)	Numeric	128 156]
49	WBC	White blood cell (cells/mL)	Numeric	[3700 18000]
50	Lymph	Lymph (lymphocyte %)	Numeric	[7 60]
51	Neut	Neutrophil (%)	Numeric	[32 89]
52	PLT	Platelet (1000/mL)	Numeric	[25 742]
53	EF	Ejection fraction(%)	Numeric	[15 60]
54	RRWMA	Region with RWMA	Numeric	[0 4]
55	VHD	Valvular heart disease{0=Normal, 1=Mild, 2=Moderate, 3=Severe}	Nominal	[0 3]
56	CATH	Target class: Cath {CD, Normal}	Binary	[0 1]

Pseudocode 1: Data Balancing

1. For each $x \in A$, where A is minority class, calculate the Euclidean distance between x and every other sample in A to find k-nearest neighbours of x.
2. Set sampling rate N according to the imbalanced proportion. For each $x \in A$, N examples (i.e x_1, x_2, …x_n) are chosen at random from its k-nearest neighbours, and they form the set A_1.
3. For each $x_k \in A_1$ $(k = 1,2,3 \ldots\ldots N)$, generate new examples by using the following formula:
$$x' = x + rand(0,1) * |x - x_k|$$
Where, *rand*(0, 1) indicates a number between 0 and 1 at random.

priate for a specific dataset. In this study, we present an Optimal Feature Subset Selection Algorithm (OFSSA) for choosing the best optimal features that utilize the characteristics of different filter-based algorithms. First, we create a rank matrix named rank_mat using algorithm 2, after getting the rank of all features using different ranking algorithms considering info gain, chi-square test, correlation-based feature selection algorithm (CFS), and relief algorithm in this study. The demonstration of Rank_mat for a dataset consisting of 8 features with 4 feature selection methods is shown in Table 3.

Table 3
Rank_mat with 4 feature selection methods for 8 features.

FSM/Feature	F1	F2	F3	F4	F5	F6	F7	F8
FSM1	6	7	3	2	5	4	8	1
FSM2	8	3	2	1	7	4	5	6
FSM3	3	8	2	1	7	5	4	6
FSM4	3	8	1	2	7	4	5	6

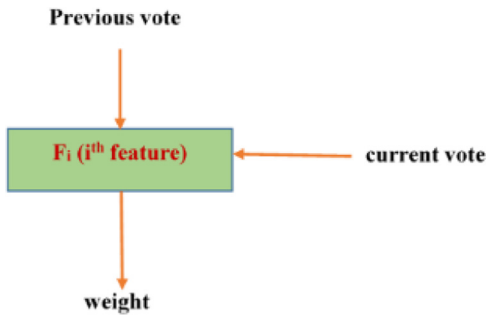


Fig. 1. Voting for one feature.

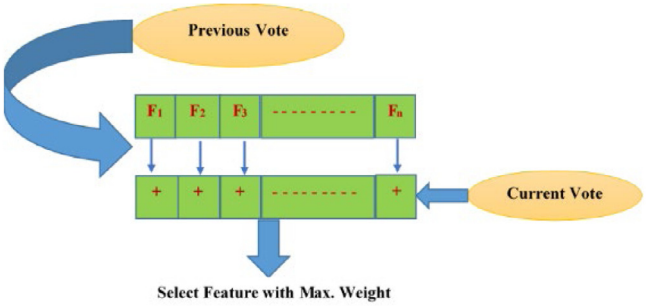


Fig. 2. Feature selection based on voting.

Algorithm 2: FRM //Feature Rank Matrix
Input: *Training Data* D^{tr}
Feature Selection Matrix: FSM
Output: *Rank Matrix rank_mat*
begin
rank_mat = Ø;
for each FSM_i = { ReliefF, Info_gain, ReliefF, chi Square, CFS}
$rank\_mat_i = FSM_i(D^{tr})$;
endfor
return rank_mat;
end

After creating a rank matrix, the OFSSA method works on the voting principle in which the selection of a feature depends upon the previous vote and the current vote as shown in Fig. 1. The previous vote is in terms of frequency of the previous rank and the current vote is the frequency of the current rank of the *i*th feature. Further, Fig. 2 demonstrates the process of voting-based feature selection, in which the previous rank is added to the current rank to get a new rank. After this, the feature with the highest rank will be selected as a new feature in the feature subset. The feature is assigned a status of −1 after selection, indicating that it will not be taken into account again during the procedure. If more than one features have the same weight value, then the selection is based on a first come first serve basis.

Further, the demonstration of feature subset selection based on the voting principle with 4 feature selection methods(FSM) and a dataset having 8 features as shown in Table 3 is presented in Table 4. Where FSM_i stands for *i*th feature selection method. FR_i is *i*th rank frequency, it is a vector consisting of the frequency of rank i for features. $FVec_i$ is a feature vector in which the selected feature gets a value −1 and other features get a value the same as FR_i. Once a feature has been chosen, it will no longer be considered in further steps.

The pseudocode of the OFSSA algorithm is given in algorithm 3. As shown in Table 4, at each *i*th iteration, OFSSA adds one new feature with maximum weight to get the *i*th feature subset FS^i. after that it will compute the performance matrix of the new feature subset. If the performance of feature subset FS^i is better than the performance of the previous feature subset FS^{i-1} then it will keep the feature subset FS^i, otherwise, it will discard FS^i and continue with FS^{i-1} for another feature.

4.2. ETCD: Effective technique for cardiac disease prediction

The goal of developing ETCD (An effective machine learning based technique for cardiac disease prediction) is to improve the accuracy, precision, and early diagnosis of cardiac disease to raise patient survival rates. The intended goal is to develop an automated solution that will aid doctors and radiologists in prognostication and decision-making with more precision and confidence along with reducing analysis time. Data pre-processing becomes a necessary step because of the high variability of a medical dataset. The framework includes a collection of datasets followed by data preprocessing (consisting of handling missing values, data normalization, and balancing). The data imbalance is dealt with using SMOTE (synthetic minority oversampling technique). After that features are extracted using OFSSA algorithm, which utilizes the strengths of several feature selection methods. The extracted features are used to train classification algorithms for normal patients and cardiac patients. The 10-fold cv partitioning is used to get the training dataset (D^{Tr}) and validation dataset (D^V). Then, ETCD will compute the average performance of all classifiers for each feature subset FS^i and returns the optimal feature subset with the best performance for all classifiers. After training, the model is validated using a validation dataset (D^v). The workflow of the proposed ETCD is depicted in detail in Fig. 3 and the pseudo-code of ETCD is given in Algorithm 4.

Algorithm 4: ETCD // Effective technique for Cardiac Disease Prediction
Input: *Dataset D*
Output: *Optimal feature subset:* FS^F
Performance matrix: P
Begin
Performance metric P = {};
D ← Data_Imputation(D);
D ← Data_Balancing(D);
(D^{Tr}, D^V) ← cv_partitioning(D);
D^{Tr} ← Data_Normalization(D^{Tr});
FSM = { ReliefF, Info_gain, chi Square, CFS}
Rank_mat = FRM(D^{tr}, FSM);
ML_Algo= { SVM, kNN, DT , NB , RF };
{FS^F, P } = **OFSSA**(D^{tr} , rank_mat, ML_Algo, P)
D^V ← Data_Normalization(D^V);
For selected features FS^F and ML_Algo
Validate D^V with FS^F for each ML_Algo
P{ Acc, Sens, Spec, precision, Score } ← ETCD_MLBox(D^V, FS^F , ML^{algo});
endfor
return (FS^F, P);
End

5. Result and discussions

The simulations of the proposed ETCD framework were carried out in MATLAB 2019 and executed on a laptop having an intel(R) Core(TM) i5 7200U @ 2.70 GHz processor with 8 GB RAM and 250 MB SSD.

5.1. Evaluation criteria

The confusion matrix is computed to evaluate the performance of the suggested framework ETCD. The primary components of the confusion matrix are True positives (TP means method correctly identified as having the cardiac disease), true negatives (TN

Table 4
Demonstration of feature subset selection based on the rank matrix.

FSM_i(Feature Selection Method) / Feature	F1	F2	F3	F4	F5	F6	F7	F8	Selected Feature subsets // Remarks
FSM1	6	7	3	2	5	4	8	1	Rank matrix
FSM2	8	3	2	1	7	4	5	6	
FSM3	3	8	2	1	7	5	4	6	
FSM4	3	8	1	2	7	4	5	6	
FR1	0	0	1	2	0	0	0	1	Frequency of feature with rank 1
FVec1	0	0	1	-1	0	0	0	1	**FS1={F4}** // F4 selected and assigned -1.
FR2	0	0	2/(1+2=3)	-1	0	0	0	0/ (1+0=1)	Frequency of feature with rank 2 and 1. Ignore the previously selected feature
FVec2	0	0	-1	-1	0	0	0	1	**FS2={ F4, F3}**// F3 selected and assigned -1.
FR3	2	1	-1	-1	0	0	0	0/ (1+0=1)	Frequency of feature with rank 3,2and 1. Ignore the previously selected feature
FVec3	-1	1	-1	-1	0	0	0	1	**FS3={ F4, F3, F1}**// F1 selected and assigned -1.
FR4	-1	0/(0+1=1)	-1	-1	0	3	1/ (0+1=1)	0/ (1+0=1)	Frequency of feature with rank 4,3,2 and 1. Ignore the previously selected feature.
FVec4	-1	1	-1	-1	0	-1	1	1	**FS4={F4, F3, F1, F6}**// F6 selected and assigned -1.
FR5	-1	0/(0+1=1)	-1	-1	1	-1	2/ (2+1=3)	0/ (0+1=1)	Frequency of feature with rank 5,4,3,2 and 1. Ignore the previously selected feature.
FVec5	-1	1	-1	-1	1	-1	-1	1	**FS5={F4, F3, F1, F6, F7}**// F7 selected and assigned -1.
FR6	-1	0/(0+1=1)	-1	-1	0/(0+1=1)	-1	-1	3/ (3+1=4)	Frequency of feature with rank 6,5,4,3,2 and 1. Ignore the previously selected feature.
FVec6	-1	1	-1	-1	1	-1	-1	-1	**FS6={ F4, F3, F1, F6, F7, F8}**// F8 selected and assigned -1.
FR7	-1	1/(1+1=2)	-1	-1	3/(3+1=4)	-1	-1	-1	Frequency of feature with rank 7,6,5,4,3,2 and 1. Ignore the previously selected feature.
FVec7	-1	2	-1	-1	-1	-1	-1	-1	**FS7={ F4, F3, F1, F6, F7, F8, F5}**// F5 selected and assigned -1.
FR8	-1	2/(2+2=4)	-1	-1	-1	-1	-1	-1	Frequency of feature with rank 8,7,6,5,4,3,2 and 1. Ignore the previously selected feature.
FVec8	-1	-1	-1	-1	-1	-1	-1	-1	**FS8={F4, F3, F1, F6, F7, F8, F5, F2}**// F2 selected and assigned -1.

Algorithm 3: OFSSA // Optimal Feature Subset Selection Algorithm

Input: *Training Dataset: D^{tr}*
Rank Matrix rank_mat
ML algorithms: ML_Algo
Performance matrix: P,
Output: *Optimal feature subset: FS^F*
Performance matrix: P

```
Begin
    RF (1: #features) =0;
    FS^0 = [Ø];
    Best_P=0;
    for F = 1: #features
        for i = 1: #features
            if RF(i) ≠ -1
                for j=1: #FSM
                    if rank_mat(j,i)==F
                        RF(i)=RF(i)+1;
                    endif
                endfor
            endif
        endfor
        [M, F_index]=max(FVec);
        FS^F = FS^(F-1) ∪ F_index;
        RF(F_index)=-1;
        D^Tr_F =  Select data according to FS^F.
        Algo^best_P = {};
        for  Valgo Є ML_Algo{SVM, kNN, DT , NB , RF}
            do
                P{ Acc, Sens, Spec, precision, Score }^algo ← ETCD_MLBox(D^Tr, FS^F , algo);
            end do
            P^avg←average(P{Acc, Sens, Spec, precision, Score });
            If Best_P < P^avg)
                Best_P = P^avg;
            else
                FS^F = FS^F -  F_index;
            endif
        endfor
    endfor
    return(FS^F, Best_P);
end
```

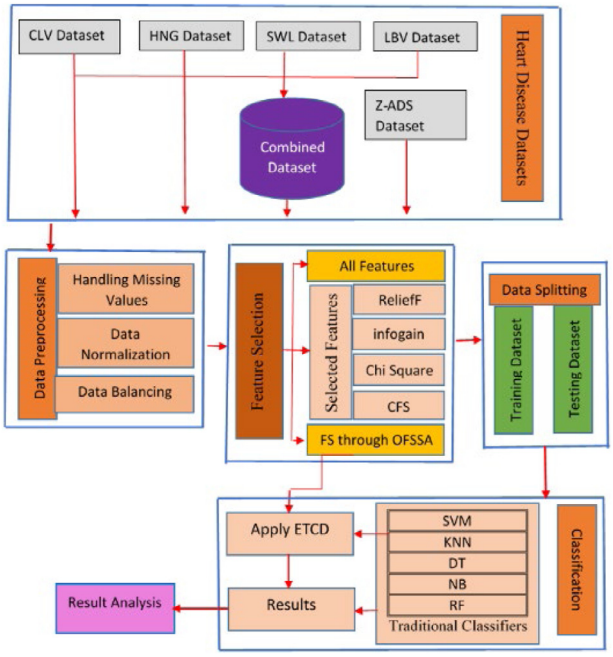


Fig. 3. Proposed framework of ETCD.

means method correctly identified the persons truly having no cardiac disease), false positives (FP means the method identifying non-cardiac disease patients as CD patients), and false negatives (FN means method identified CD patients as normal patients). The presented framework ETCD follows the matric (Acc, Sens, Spec, precision, F_Score) for cardiac disease diagnostics. The accuracy (Acc) is defined as the percentage ratio of correctly identified patients to the total number of patients in a given class. Only Acc was unable to make a precise distinction between cardiac patients and normal persons. The cardiac and normal persons are classified as Sensitivity(Sens) and specificity(Spec). Whereas, precision defines the fraction of relevant results among all retrieved results. The score is a weighted average of recall and precision that can be expressed quantitatively. The metric (Acc, Sens, Spec, precision, F_Score) is computed by using Eq. (2)–(6).

$$Acc = \frac{correctly\ classified\ patients}{total\ number\ of\ patients} * 100\% \tag{2}$$

Where correctly classified patients are computed as (TN + TP) and the total number of patients is calculated as (TP + FP + FN + TN).

$$Sens = \frac{TP}{TP + FN} * 100\% \tag{3}$$

$$Spec = \frac{TN}{FP + TN} * 100\% \tag{4}$$

$$Precision = \frac{TP}{TP + FP} * 100\% \tag{5}$$

$$F\_Score = \frac{2 * TP}{2 * TP + FP + FN} * 100\% \tag{6}$$

5.2. Performance evaluation

5.2.1. Performance evaluation with and without balancing

5.2.1.1. Accuracy comparison w.r.t. datasets. The performance of datasets originally and with Balancing considering all features is presented in Table 5. This table compares the performance of the considered datasets for SVM, KNN, DT, NB, and RF classifiers. For Cleveland and Hungarian datasets, all classifiers perform better after balancing as compared to the original dataset (Table 5).

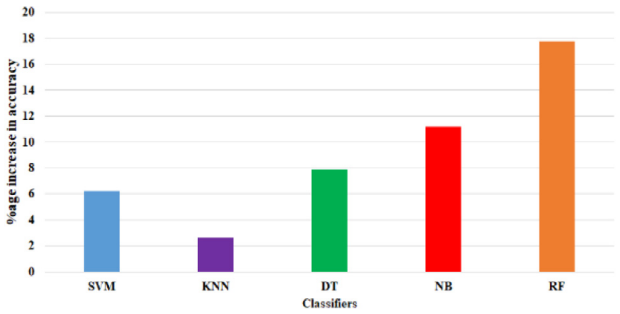


Fig. 4. Average percentage increase in accuracy for classifiers.

The average performance accuracy for the Cleveland dataset gets increases by 8.94% after balancing and for the Hungarian dataset, the average accuracy gets increases by 8.74% after balancing. For the combined dataset, after balancing the accuracy increased by 5.94% for SVM, KNN, and DT classifier, but for NB and RF classifiers, accuracy is decreased by 5.6% after balancing. For the Z_Alizadeh dataset, the accuracy with all classifiers gets increased after balancing. With balancing, classifiers give 13.004% better accuracy for the Z_Alizadeh dataset.

5.2.1.2. Accuracy comparison w.r.t. classifiers. The performance of different classifiers has been compared for datasets with and without balancing. Fig. 4 shows that the performance accuracy of SVM, KNN, DT, NB, and RF classifiers increased by 6.20%, 2.65%,7.91%, 11.19%, 17.72% respectively with balanced datasets. Further, Fig. 5 shows the reflection of execution time taken by different classifiers for considered datasets. There is an 11.2% increase in execution time after balancing.

5.2.2. Performance evaluation of different FS techniques with various classifiers

In this section, we evaluate the performance of various feature ranking methods on the above-mentioned datasets with different classifiers. The considered feature ranking methods are ReliefF, Info-Gain, Chi-Square and Correlation based feature selection. Table 6 represents the rank given to various dataset features by different feature selection methods. For evaluation, features are considered based on the 'top N features' strategy. Where 'N' is 50%, 60%, and 70% in this study.

5.2.2.1. Performance evaluation of different FS techniques with various classifiers for Cleveland dataset. Table 7 shows the performance of various ranking methods with different classifiers for the Cleveland dataset. With the ReliefF FS method, SVM and KNN provide better accuracy of 91.42% and 97.69% respectively with top 50% features. The DT and RF perform well with top 70% features with 90.10% and 80.53% accuracy. Whereas, NB gives better accuracy 81.19% with top 60% features. On the other side, features selected by the Info gain method, SVM, DT, NB, RF produce the best accuracy of 91.09%, 91.09%, 79,54%, and 78.55% respectively with top 70% features and KNN gives the best accuracy i.e. 95,05% with top 50% and 60% features. Whereas, with the chi-square technique, SVM gives the best performance accuracy of 92.08% with the top 60% features. KNN, DT, and NB provide the best accuracy of 90.76%, 90.76%, and 80.86% respectively with top 60% features and RF performs well with top 50% features with 81.85% accuracy. For the correlation-based FS technique, SVM gives an accuracy 89.44% with the top 50% and 70% of the features. KNN gives the best accuracy 97.03% with top 50% and 60% features and DT, NB, RF classifiers produce better accuracy with top 70% features, that is, 89.44%, 78.22%, and 70.63% resp.

Table 5
Performance evaluation of Original dataset Vs Balanced dataset with all features.

Dataset	Classifiers	Original dataset					Balanced dataset (SMOTE)				
		Accuracy	Sensitivity	Specificity	Precision	F_Score	Accuracy	Sensitivity	Specificity	Precision	F_Score
Cleveland	SVM	92.739	96.063	90.341	87.770	91.729	**99.545**	99.641	99.029	99.820	99.730
	KNN	97.030	98.507	95.858	94.964	96.703	**99.848**	100.000	99.048	99.820	99.910
	DT	88.779	88.889	88.690	86.331	87.591	**98.333**	99.277	93.458	98.741	99.008
	NB	81.848	86.207	79.144	71.942	78.431	**89.848**	93.583	68.687	94.424	94.002
	RF	79.868	97.561	73.303	57.554	72.398	**95.606**	96.803	88.660	98.022	97.408
Hungarian	SVM	92.177	92.784	91.878	84.906	88.670	**97.491**	97.892	96.183	98.818	98.353
	KNN	95.918	92.727	97.826	96.226	94.444	**99.499**	99.499	98.526	100.000	99.749
	DT	92.177	90.291	93.194	87.736	88.995	**98.029**	98.131	97.692	99.291	98.707
	NB	**83.333**	84.337	82.938	66.038	74.074	75.986	76.225	57.143	99.291	86.242
	RF	74.490	91.892	71.984	32.075	47.552	**92.294**	92.601	91.071	97.636	95.052
Combined dataset	SVM	90.280	95.364	88.596	73.469	82.997	**97.799**	98.346	96.139	98.723	98.534
	KNN	97.445	96.196	98.015	95.676	95.935	**99.666**	99.832	98.99	99.832	99.832
	DT	89.624	86.093	90.931	77.844	81.761	**97.318**	98.458	93.985	97.954	98.205
	NB	**76.373**	73.585	76.970	40.625	52.349	75.024	75.000	100.000	100.000	85.714
	RF	**69.558**	100.000	69.454	41.105	62.186	62.967	100.000	47.370	50.575	67.176
Z_alizadeh	SVM	94.719	93.860	97.333	99.074	96.396	**99.556**	99.539	100.000	100.000	99.769
	KNN	97.690	97.717	97.619	99.074	98.391	**99.666**	99.832	100.000	99.832	99.832
	DT	91.419	93.578	85.882	94.444	94.009	**99.445**	99.884	90.000	99.537	99.710
	NB	72.277	72.000	85.000	85.000	83.721	**95.671**	95.987	75.000	99.653	97.785
	RF	71.287	71.287	72.365	84.000	83.237	**95.893**	95.893	94.256	92.000	97.904

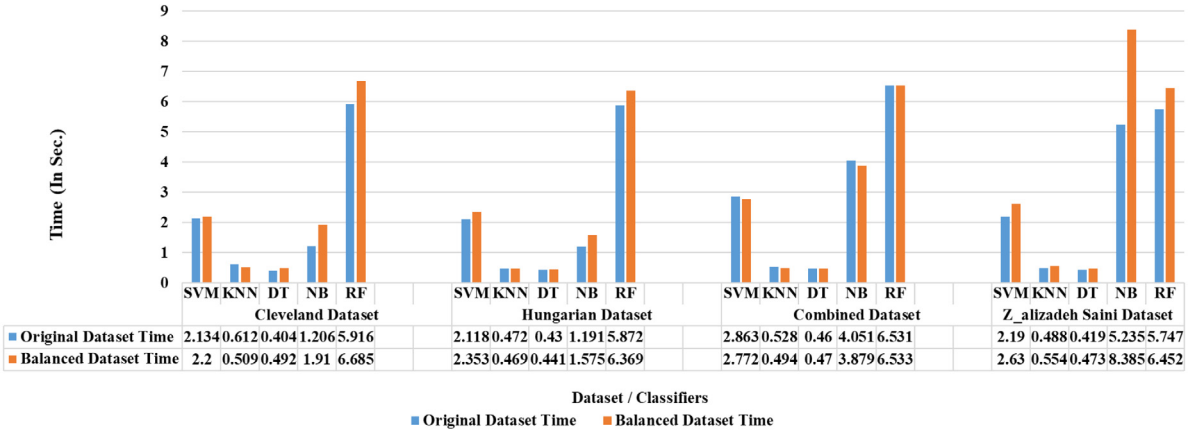


Fig. 5. Time taken by classifiers for various datasets with and without balancing.

5.2.2.2. Performance evaluation of different FS techniques with various classifiers for Hungarian dataset. The performance of various ranking methods with different classifiers for the Hungarian dataset is presented in Table 8. For the ReliefF FS method, SVM, KNN, and DT provide better accuracy of 92.86%, 97.28%, and 89.80% with top 70% features. The NB gives better accuracy 82.31% with top 60% features and RF performs well with top 50% and 70% features with 64.97% accuracy. On another side, with the Info gain method, SVM gives better accuracy 91.50% with top 50% and 70% features. KNN and NB produce the best accuracy 96.94% and 80.95% with top 70% features. DT gives the best accuracy of 90.48% with top 60% and 70% features and RF gives the best accuracy 71.09% with top 50% and 60% features. On the other side, for the chi-square method, SVM, DT, NB, and RF classifiers give the best performance accuracy of 91.84%, 90.48%, 82.31%, and 73.81% with top 70% features. While KNN gives 96.26% accuracy with the top 60% features. For the correlation-based FS technique, SVM, KNN, DT, NB, and RF all classifiers give the best accuracy with 70% of the features and that is 92.86%, 97.62%, 90.14%, 81.97%, and 64.63% respectively.

5.2.2.3. Performance evaluation of different FS techniques with various classifiers for Combined dataset. Table 9 shows the results of various ranking techniques with various classifiers for the Combined dataset. For ReliefF FS method, SVM, KNN, DT and RF provide better accuracy of 89.95%, 96.93%, 87.84% and 73.32% with top 70% features. The NB gives better accuracy 75.47% with top 50% features. On other hand, for the Info gain method, SVM and DT produce 89.95% and 88.93% accuracy with the top 70% features. KNN and NB produce the best accuracy of 98.27% and 73.47% with top 60% features and RF gives the best accuracy of 70.54% with top 50% features. For the chi-square method, SVM, DT, and NB classifiers produce the best accuracy of 88.80%, 89.23%, and 46.04% with top 70% features. While KNN and RF give 97.58% and 69.06% accuracy with top 60% features. For the correlation-based FS technique, SVM gives an accuracy 89.29% with top 70% features. KNN, DT, and RF classifiers provide better accuracy with 50% features and that is 97.29%, 89.89%, and 71.01% respectively. Whereas, NB performs better with 60% features with 77.53% accuracy.

5.2.2.4. Performance evaluation of different FS techniques with various classifiers for Z_Alizadeh Saini dataset. The performance of several ranking strategies with various classifiers for the Z_Alizadeh Saini dataset is shown in Table 10. For ReliefF FS method, SVM, KNN, DT provide better accuracy of 94.06%, 95.05%, and 91.75% with top 70% features. The NB gives better accuracy 71.62% with the top 60% features and RF produces 71.29% accuracy with all 50%, 60%, and 70% features. With the Info gain method, SVM and DT provide 93.73% and 92.41% accuracy with

Table 6
Feature ranking by different feature selection techniques to various datasets.

Dataset	Feature selection technique	Feature ranking
Cleveland dataset	ReliefF	3, 12, 1, 13, 11, 6, 9, 8, 2, 4, 10, 5, 7
	Info Gain	7, 2, 9, 6, 3, 11, 5, 4, 12, 13, 1, 10, 8
	Chi Square	5, 8, 10, 13, 3, 12, 9, 1, 4, 11, 2, 7, 6
	Correlation based feature selection	5, 8, 11, 6, 10, 2, 3, 7, 13, 9, 4, 12, 1
Hungarian dataset	ReliefF	13, 5, 12, 6, 10, 7, 9, 11, 8, 3, 4, 1, 2
	Info Gain	2, 11, 13, 3, 1, 6, 5, 12, 7, 4, 8, 10, 9
	Chi Square	5, 11, 9, 10, 13, 3, 7, 1, 4, 12, 9, 2, 5
	Correlation based feature selection	6, 8, 11, 10, 13, 3, 7, 1, 4, 12, 9, 2, 5
Combined dataset	ReliefF	10, 3, 11, 12, 9, 7, 4, 13, 6, 2, 8, 5, 1
	Info Gain	13, 2, 11, 1, 6, 3, 5, 8, 4, 7, 12, 9, 10
	Chi Square	4, 5, 8, 10, 3, 9, 11, 13, 6, 2, 12, 7, 1
	Correlation based feature selection	7, 10, 4, 9, 5, 6, 8, 2, 11, 3, 13, 12, 1
Z_Alizadeh Saini dataset	ReliefF	6, 36, 25, 17, 26, 5, 29, 28, 10, 24, 16, 13, 15, 1, 2, 7, 18, 48, 9, 31, 47, 20, 3, 51, 52, 38, 41, 42, 43, 50, 22, 14, 53, 21, 35, 12, 33, 32, 49, 23, 8, 54, 39, 37, 45, 46, 44, 19, 55, 40, 34, 27, 4, 11, 30
	Info Gain	26, 25, 17, 7, 4, 11, 35, 28, 6, 34, 8, 10, 27, 23, 54, 36, 29, 55, 30, 32, 14, 20, 33, 38, 9, 15, 37, 13, 22, 18, 24, 42, 50, 31, 12, 21, 47, 5, 49, 19, 51, 46, 43, 16, 40, 48, 52, 44, 53, 45, 3, 2, 1, 41, 39
	Chi Square	5, 41, 52, 39, 42, 25, 50, 1, 46, 49, 43, 28, 45, 51, 53, 2, 3, 54, 44, 47, 18, 7, 19, 29, 48, 6, 40, 35, 55, 32, 24, 34, 33, 27, 26, 37, 12, 38, 14, 21, 8, 4, 20, 31, 36, 15, 22, 11, 16, 10, 9, 13, 17, 23, 30
	Correlation based feature selection	37, 55, 7, 54, 18, 11, 47, 27, 12, 17, 40, 46, 49, 43, 52, 42, 14, 5, 21, 29, 45, 23, 50, 22, 31, 19, 36, 38, 51, 13, 48, 8, 26, 2, 25, 15, 6, 1, 32, 39, 35, 3, 20, 9, 10, 16, 33, 24, 34, 4, 41, 44, 28, 53, 30

Table 7
Performance evaluation of different FS techniques with Top N features for Cleveland dataset.

Feature selection technique	Classifier	Top 50% features					Top 60% features					Top 70% features				
		Accuracy	Sensitivity	Specificity	Precision	F_Score	Accuracy	Sensitivity	Specificity	Precision	F_Score	Accuracy	Sensitivity	Specificity	Precision	F_Score
ReliefF+ top N Features	SVM	**91.42**	93.80	89.66	87.05	90.30	90.10	91.60	88.95	86.33	88.89	90.76	91.73	90.00	87.77	89.71
	KNN	**97.69**	97.83	97.58	97.12	97.47	96.37	97.06	95.81	94.96	96.00	96.04	97.04	95.24	94.24	95.62
	DT	87.79	86.96	88.48	86.33	86.64	88.45	88.24	88.62	86.33	87.27	**90.10**	91.60	88.95	86.33	88.89
	NB	80.20	83.76	77.96	70.50	76.56	**81.19**	84.75	78.92	71.94	77.82	80.86	82.93	79.44	73.38	77.86
	RF	78.22	83.49	75.26	65.47	73.39	80.20	88.35	76.00	65.47	75.21	**80.53**	94.44	74.65	61.15	74.24
Info Gain + top N Features	SVM	85.48	89.26	82.97	77.70	83.08	89.77	92.86	87.57	84.17	88.30	**91.09**	95.16	88.27	84.89	89.73
	KNN	**95.05**	96.27	94.08	92.81	94.51	**95.05**	96.27	94.08	92.81	94.51	94.39	95.52	93.49	92.09	93.77
	DT	87.13	88.46	86.13	82.73	85.50	86.47	87.12	85.96	82.73	84.87	**91.09**	92.42	90.06	87.77	90.04
	NB	74.92	76.47	73.91	65.47	70.54	77.23	81.82	74.61	64.75	72.29	**79.54**	80.31	78.98	73.38	76.69
	RF	66.34	89.36	62.11	30.22	45.16	66.01	97.37	61.51	26.62	41.81	**78.55**	84.91	75.13	64.75	73.47
Chi-Square + top N Features	SVM	91.75	95.97	88.83	85.61	90.49	**92.08**	94.57	90.23	87.77	91.04	91.09	93.08	89.60	87.05	89.96
	KNN	96.37	98.48	94.74	93.53	95.94	96.37	98.48	94.74	93.53	95.94	**96.70**	99.24	94.77	93.53	96.30
	DT	87.79	89.84	86.29	82.73	86.14	88.45	91.27	86.44	82.73	86.79	**90.76**	89.93	91.46	89.93	89.93
	NB	78.55	77.21	79.64	75.54	76.36	80.20	78.83	81.33	77.70	78.26	**80.86**	82.40	79.78	74.10	78.03
	RF	**81.85**	87.50	78.53	70.50	78.09	77.23	96.05	70.93	52.52	67.91	74.59	98.44	68.20	45.32	62.07
Correlation-based feature selection + top N Features	SVM	**89.44**	89.63	89.29	87.05	88.32	88.12	85.52	90.51	89.21	87.32	**89.44**	87.41	91.25	89.93	88.65
	KNN	**97.03**	96.43	97.55	97.12	96.77	**97.03**	96.43	97.55	97.12	96.77	94.39	95.52	93.49	92.09	93.77
	DT	88.12	87.59	88.55	86.33	86.96	85.81	86.36	85.38	82.01	84.13	**89.44**	86.90	91.77	90.65	88.73
	NB	74.26	72.93	75.29	69.78	71.32	77.56	77.52	77.59	71.94	74.63	**78.22**	81.20	76.34	68.35	74.22
	RF	67.00	64.89	68.60	61.15	62.96	66.01	86.00	62.06	30.94	45.50	**70.63**	94.64	65.18	38.13	54.36

Table 8
Performance evaluation of different FS techniques with Top N features for Hungarian dataset.

Feature selection technique	Classifier	Top 50% features					Top 60% features					Top 70% features				
		Accuracy	Sensitivity	Specificity	Precision	F_Score	Accuracy	Sensitivity	Specificity	Precision	F_Score	Accuracy	Sensitivity	Specificity	Precision	F_Score
ReliefF+ top N Features	SVM	88.10	91.76	86.60	73.58	81.68	92.18	97.70	89.86	80.19	88.08	**92.86**	94.74	91.96	84.91	89.55
	KNN	88.10	79.34	94.22	90.57	84.58	91.84	85.34	96.07	93.40	89.19	**97.28**	96.23	97.87	96.23	96.23
	DT	86.73	86.02	87.06	75.47	80.40	88.78	91.01	87.80	76.42	83.08	**89.80**	83.33	93.89	89.62	86.36
	NB	77.89	85.96	75.95	46.23	60.12	**82.31**	88.57	80.36	58.49	70.45	81.97	85.33	80.82	60.38	70.72
	RF	**64.97**	100.00	64.60	22.83	35.50	64.29	100.00	64.16	20.94	31.87	**64.97**	80.00	64.71	23.77	37.21
Info Gain + top N Features	SVM	**91.50**	96.55	89.37	79.25	87.05	89.80	91.30	89.11	79.25	84.85	**91.50**	92.63	90.95	83.02	87.56
	KNN	95.92	91.96	98.35	97.17	94.50	96.60	92.86	98.90	98.11	95.41	**96.94**	93.69	98.91	98.11	95.85
	DT	86.39	85.87	86.63	74.53	79.80	**90.48**	85.45	93.48	88.68	87.04	**90.48**	85.45	93.48	88.68	87.04
	NB	78.57	87.72	76.37	47.17	61.35	80.27	87.50	78.26	52.83	65.88	**80.95**	83.78	80.00	58.49	68.89
	RF	**71.09**	80.00	69.88	26.42	39.72	**71.09**	75.61	70.36	29.25	42.18	70.75	85.71	69.17	22.64	35.82
Chi-Square + top N Features	SVM	90.14	92.31	89.16	79.25	85.28	89.12	89.36	89.00	79.25	84.00	**91.84**	92.71	91.41	83.96	88.12
	KNN	94.90	90.27	97.79	96.23	93.15	**96.26**	92.79	98.36	97.17	94.93	95.92	93.52	97.31	95.28	94.39
	DT	87.76	87.23	88.00	77.36	82.00	90.14	92.31	89.16	79.25	85.28	**90.48**	88.24	91.67	84.91	86.54
	NB	81.29	82.28	80.93	61.32	70.27	80.61	81.01	80.47	60.38	69.19	**82.31**	82.93	82.08	64.15	72.34
	RF	70.75	85.71	69.17	22.64	35.82	70.75	95.45	68.75	19.81	32.81	**73.81**	96.77	71.10	28.30	43.80
Correlation-based feature selection + top N Features	SVM	89.12	91.11	88.24	77.36	83.67	89.80	92.22	88.73	78.30	84.69	**92.86**	96.70	91.13	83.02	89.34
	KNN	94.90	92.52	96.26	93.40	92.96	95.58	93.46	96.79	94.34	93.90	**97.62**	96.26	98.40	97.17	96.71
	DT	85.71	85.56	85.78	72.64	78.57	87.76	88.04	87.62	76.42	81.82	**90.14**	85.98	92.51	86.79	86.38
	NB	76.53	81.36	75.32	45.28	58.18	77.21	84.21	75.53	45.28	58.90	**81.97**	87.32	80.27	58.49	70.06
	RF	64.29	60.00	64.36	42.83	45.41	64.29	66.67	64.26	41.89	43.67	**64.63**	75.00	64.48	42.83	45.45

Table 9
Performance evaluation of different FS techniques with Top N features for Combined dataset.

Feature selection technique	Classifier	Top 50% features					Top 60% features					Top 70% features				
		Accuracy	Sensitivity	Specificity	Precision	F_Score	Accuracy	Sensitivity	Specificity	Precision	F_Score	Accuracy	Sensitivity	Specificity	Precision	F_Score
ReliefF+ top N Features	SVM	89.13	95.77	87.10	69.39	80.47	89.46	97.14	87.15	69.39	80.95	**89.95**	91.41	89.41	76.02	83.01
	KNN	95.21	90.26	97.69	95.14	92.63	95.40	90.77	97.70	95.16	92.91	**96.93**	94.12	98.25	96.17	95.14
	DT	88.24	84.42	89.68	75.58	79.75	88.85	89.47	88.65	71.26	79.33	**87.84**	87.88	87.82	69.05	77.33
	NB	**75.47**	66.99	77.29	38.76	49.11	66.01	48.38	84.23	76.02	59.13	72.16	55.13	85.17	73.98	63.18
	RF	71.94	77.78	71.75	38.43	35.22	73.06	90.00	72.74	35.84	33.98	**73.32**	100.00	73.27	30.69	31.38
Info Gain + top N Features	SVM	87.48	91.10	86.33	67.86	77.78	87.48	86.14	87.98	72.96	79.01	**89.95**	90.91	89.59	76.53	83.10
	KNN	97.93	98.24	97.80	94.89	96.53	**98.27**	98.26	98.28	96.02	97.13	97.98	97.33	98.28	96.30	96.81
	DT	88.38	89.17	88.17	67.72	76.98	87.99	84.38	89.10	70.59	76.87	**88.93**	87.23	89.50	73.65	79.87
	NB	73.13	79.63	72.50	22.16	34.68	**73.47**	79.31	72.84	23.71	36.51	69.50	52.50	78.00	54.40	53.44
	RF	**70.54**	60.00	70.54	20.00	22.08	68.78	65.50	68.78	20.50	25.26	69.15	67.07	69.10	23.55	21.09
Chi-Square + top N Features	SVM	88.14	86.47	88.79	75.00	80.33	86.66	85.28	87.16	70.92	77.44	**88.80**	91.03	88.03	72.45	80.68
	KNN	95.50	91.98	97.19	93.99	92.97	**97.58**	96.13	98.24	96.13	96.13	97.30	95.74	98.02	95.74	95.74
	DT	88.44	82.53	90.86	78.74	80.59	88.99	87.41	89.52	73.96	80.13	**89.23**	88.97	89.31	72.89	80.13
	NB	33.77	32.60	80.00	98.47	48.98	38.06	33.99	88.89	97.45	50.40	**46.04**	37.13	92.78	96.43	53.62
	RF	68.49	100.00	68.33	21.55	23.06	**69.06**	100.00	68.96	31.07	22.12	69.00	100.00	68.84	21.59	23.13
Correlation-based feature selection + top N Features	SVM	88.30	86.98	88.81	75.00	80.55	89.13	90.63	88.59	73.98	81.46	**89.29**	92.81	88.11	72.45	81.38
	KNN	**97.92**	95.63	98.99	97.77	96.69	98.45	97.83	98.74	97.30	97.56	98.47	98.36	98.52	96.77	97.56
	DT	**89.89**	89.71	89.95	74.85	81.61	89.48	85.16	91.13	78.57	81.73	89.21	81.93	92.31	81.93	81.93
	NB	63.45	42.33	91.23	86.39	56.82	**77.53**	47.76	82.80	32.99	39.02	77.00	66.22	78.76	33.79	44.75
	RF	**71.01**	81.25	70.71	37.34	33.47	69.44	87.50	69.19	33.76	37.22	69.23	87.50	68.98	33.68	37.07

Table 10
Performance evaluation of different FS techniques with Top N features for Z_Alizadeh Saini dataset.

Feature selection technique	Classifier	Top 50% features					Top 60% features					Top 70% features				
		Accuracy	Sensitivity	Specificity	Precision	F_Score	Accuracy	Sensitivity	Specificity	Precision	F_Score	Accuracy	Sensitivity	Specificity	Precision	F_Score
ReliefF+ top N Features	SVM	91.75	90.30	96.97	99.07	94.48	93.40	92.98	94.67	98.15	95.50	**94.06**	93.42	96.00	98.61	95.95
	KNN	94.39	94.62	93.75	97.69	96.13	94.39	94.62	93.75	97.69	96.13	**95.05**	94.67	96.15	98.61	96.60
	DT	89.77	91.86	84.15	93.98	92.91	91.09	92.76	86.59	94.91	93.82	**91.75**	93.21	87.80	95.37	94.28
	NB	71.95	71.76	98.75	96.00	83.56	**71.62**	71.52	97.63	96.00	83.40	71.95	71.76	98.56	96.00	83.56
	RF	**71.29**	71.29	70.99	96.00	83.24	**71.29**	71.29	71.43	96.00	83.24	**71.29**	71.29	75.73	96.00	83.24
Info Gain + top N Features	SVM	91.09	90.56	92.86	97.69	93.99	89.77	89.03	92.42	97.69	93.16	**93.73**	93.39	94.74	98.15	95.71
	KNN	**95.05**	95.07	95.00	98.15	96.58	94.72	95.05	93.83	97.69	96.35	**95.05**	95.07	95.00	98.15	96.58
	DT	89.77	90.75	86.84	95.37	93.00	91.42	91.67	90.67	96.76	94.14	**92.41**	93.67	89.02	95.83	94.74
	NB	**72.28**	72.00	96.00	96.00	83.72	**72.28**	72.00	96.35	96.00	83.72	**72.28**	72.00	97.63	96.60	83.72
	RF	**71.29**	71.29	68.81	96.00	83.24	**71.29**	71.29	66.97	96.00	83.24	**71.29**	71.29	69.72	96.00	83.24
Chi-Square + top N Features	SVM	89.77	88.70	93.75	98.15	93.19	90.10	88.43	96.72	99.07	93.45	**91.42**	91.30	91.78	97.22	94.17
	KNN	94.72	94.64	94.94	98.15	96.36	94.72	94.64	94.94	98.15	96.36	**95.05**	95.07	95.00	98.15	96.58
	DT	**91.75**	92.07	90.79	96.76	94.36	88.45	88.51	88.24	96.30	92.24	91.42	92.41	88.61	95.83	94.09
	NB	**72.28**	72.00	99.23	96.00	83.72	**72.28**	72.00	95.623	96.00	83.72	**72.28**	72.00	98.56	96.00	83.72
	RF	71.62	71.52	69.16	96.00	83.40	**71.95**	71.76	72.90	96.00	83.56	71.29	71.29	66.98	96.00	83.24
correlation-based feature selection + top N Features	SVM	88.12	87.19	91.80	97.69	92.14	92.74	92.17	94.52	98.15	95.07	**93.73**	92.64	97.22	99.07	95.75
	KNN	**95.71**	94.32	98.95	99.2	97.08	95.38	95.09	96.20	98.61	96.82	94.72	94.64	94.94	98.15	96.36
	DT	88.12	88.79	85.92	95.37	91.96	**93.73**	94.17	92.50	97.22	95.67	92.74	95.33	86.52	94.44	94.88
	NB	71.95	71.76	98.5	96.00	83.56	**72.28**	72.00	93.025	96.50	83.72	71.95	71.76	100.00	96.00	83.56
	RF	71.29	71.29	66.99	95.00	83.24	71.29	71.29	71.70	95.43	83.24	**71.62**	71.52	100.00	96.00	83.40

70% features. KNN gives 95.05% accuracy with both 50% and 70% features. NB and RF both produce 72.28% and 71.29% accuracy with all 50%, 60%, and 70% features. For the chi-square method, SVM, and KNN produce the best accuracy of 91.42% and 95.05% with top 70% features. DT gives 91.75% accuracy with the top 50% features. NB gives the best accuracy of 72.28% with all 50%, 60%, and 70% features. RF provides 71.95% accuracy with the top 60% features. For the correlation-based FS technique, SVM and RF give 93.73% and 71.62% accuracy with the top 70% features. KNN provides better accuracy 95.71% with 50% features Whereas, DT and NB perform better with 60% features and that is 93.73% and 72.28% accuracy.

5.2.3. Performance evaluation of ETDC

The performance of the present framework ETCD is presented in Table 11. For Cleveland dataset, the best selected feature set is {3,12,1,2,5,10,11,13} for which SVM, KNN, DT, NB, and RF performs with 97.879%, 99.242%, 97.273%, 90.303, and 94.545% accuracy respectively. For Hungarian dataset, the selected feature set is {5,2,1,4,9,11,3,10}. With this feature set the SVM, KNN, DT, NB, and RF classifiers give 97.491%, 99.187%, 98.208%, 83.333%, and 91.039% accuracy respectively. For combined dataset, SVM, KNN, DT, NB, and RF classifiers performs with 97.268%, 99.287%, 98.231%, 83.445% and 87.211% accuracy for selected feature set {13,2,5,4,11,1,6,9,8}. For Z_Alizadeh Saini dataset, the selected feature subset is {2,1,7,4,5,11,12,53,9,20,13,23,15,25,52,16,32,22, 33,39,37}. For this feature subset, SVM performs with 98.668% accuracy, KNN performs with 99.666% accuracy, whereas DT, NB, and RF perform with 98.113%, 95.893%, and 95.893% accuracy respectively. Further, Fig. 6 presents the percentage of features selected for different datasets. For Cleveland and Hungarian datasets, ETCD selected 61.54% features. For the combined dataset, it chooses 69.23% and for the Z_Alizadeh dataset, the selected features are 38.18%.

5.2.3.1. Accuracy comparison of ETCD with other FS techniques. A comparison of performance accuracy achieved from ETCD with other FS methods used in this study is shown in Fig. 7. It is clear from Fig. 7 that ETCD enhances the performance accuracy of all classifiers for all datasets. Table 12 shows the percentage difference between the average accuracy of ReliefF, Info gain, Chi-square, and Correlation-based feature selection with ETCD. There is a 9.186% increase in accuracy with ETCD as compared to ReliefF and a 9.416% increase as compared to Info gain. Whereas, with ETCD, there is a 10.58% and 9.824% increase in accuracy as compared to chi-square and Correlation-based feature selection. Furthermore, the *p*-value (Table 12) obtained after performing paired t-tests on the performance of the other conventional feature ranking algorithms with ETCD is less than 0.05, indicating that the proposed ETCD has substantial performance.

5.2.3.2. Comparison of ETCD with other FS techniques for other performance metrics. Other performance metrics such as sensitivity, specificity, precision, and F_Score have also been used to evaluate the performance of various methods. Fig. 8(a) represents the sensitivity score, which is an important performance matrix because appropriately classifying persons with cardiac disease is crucial. It is clear from Fig. 8(a) that the sensitivity score of ETCD is better for all datasets as compared to other methods. With ETCD, the KNN classifier produced 100% sensitivity for Cleveland

Table 11
Performance evaluation of ETCD for various datasets with different classifiers.

Dataset	Classifier	Accuracy	Sensitivity	Specificity	Precision	F_Score	Selected features
Cleveland dataset	SVM	97.879	99.094	91.667	98.381	98.736	3, 12, 1, 2, 5, 10, 11, 13
	KNN	99.242	100.000	95.413	99.101	99.548	
	DT	97.273	98.909	89.091	97.842	98.373	
	NB	90.303	93.929	79.000	94.604	94.265	
	RF	94.545	96.763	82.692	96.763	96.763	
Hungarian dataset	SVM	97.491	98.118	95.489	98.582	98.349	5, 2, 1, 4, 9, 11, 3, 10
	KNN	99.187	100.000	99.002	99.130	99.563	
	DT	98.208	98.818	96.296	98.818	98.818	
	NB	83.333	84.375	76.923	95.745	89.701	
	RF	91.039	92.874	84.553	95.508	94.172	
Combined dataset	SVM	97.268	96.851	94.422	98.212	97.527	13, 2, 5, 4, 11, 1, 6, 9, 8
	KNN	99.287	99.138	9.533	100.000	99.567	
	DT	98.321	98.710	93.333	97.701	98.203	
	NB	83.445	82.937	87.395	98.084	89.877	
	RF	87.211	84.211	75.527	65.283	68.896	
Z_Alizadeh Saini dataset	SVM	98.668	98.630	97.5963	100.000	99.310	2, 1, 7, 4, 5, 11, 12, 53, 9, 20, 13, 23, 15, 25, 52, 16, 32, 22, 33, 39, 37
	KNN	99.666	99.832	98.636	99.832	99.832	
	DT	98.113	98.734	97.235	99.306	99.019	
	NB	95.893	95.893	98.92308	100.000	98.904	
	RF	95.893	95.893	89.237	100.000	97.904	

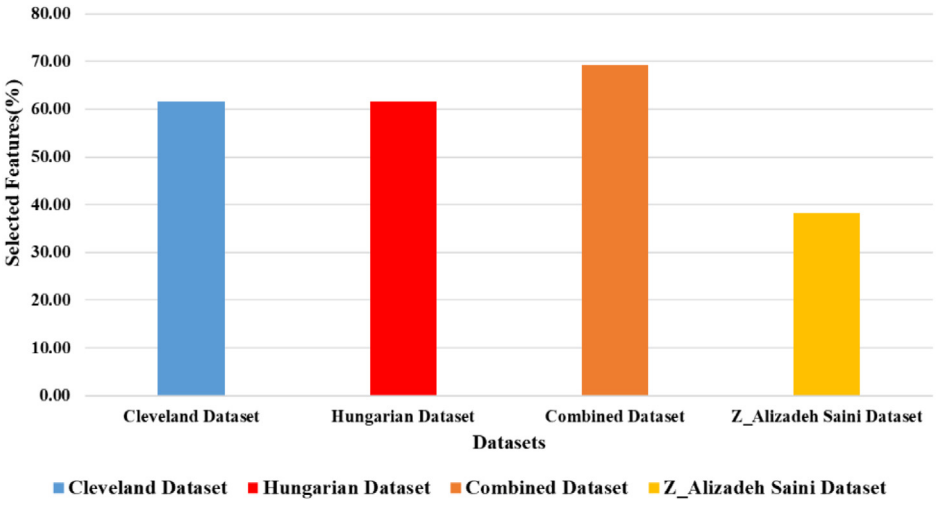


Fig. 6. Percentage of features selected by ETCD for various datasets.

Fig. 7. Performance accuracy comparison of ETCD with other FS techniques for various datasets.

Table 12
Percentage difference of average accuracy and *p*-value obtained from paired t-test of ETCD with other FS techniques.

	ReliefF + top N Features	Info Gain + top N Features	Chi-Square + top N Features	Correlation-based feature selection + top N Features	Proposed ETCD
Average accuracy	85.8215	85.605	84.5025	85.219	94.5035
%age difference with proposed ETCD	9.186961329	9.416053374	10.58267683	9.824503854	–
Paired t-test	3.50013E−05	6.67525E−06	4.65242E−05	3.86221E−05	

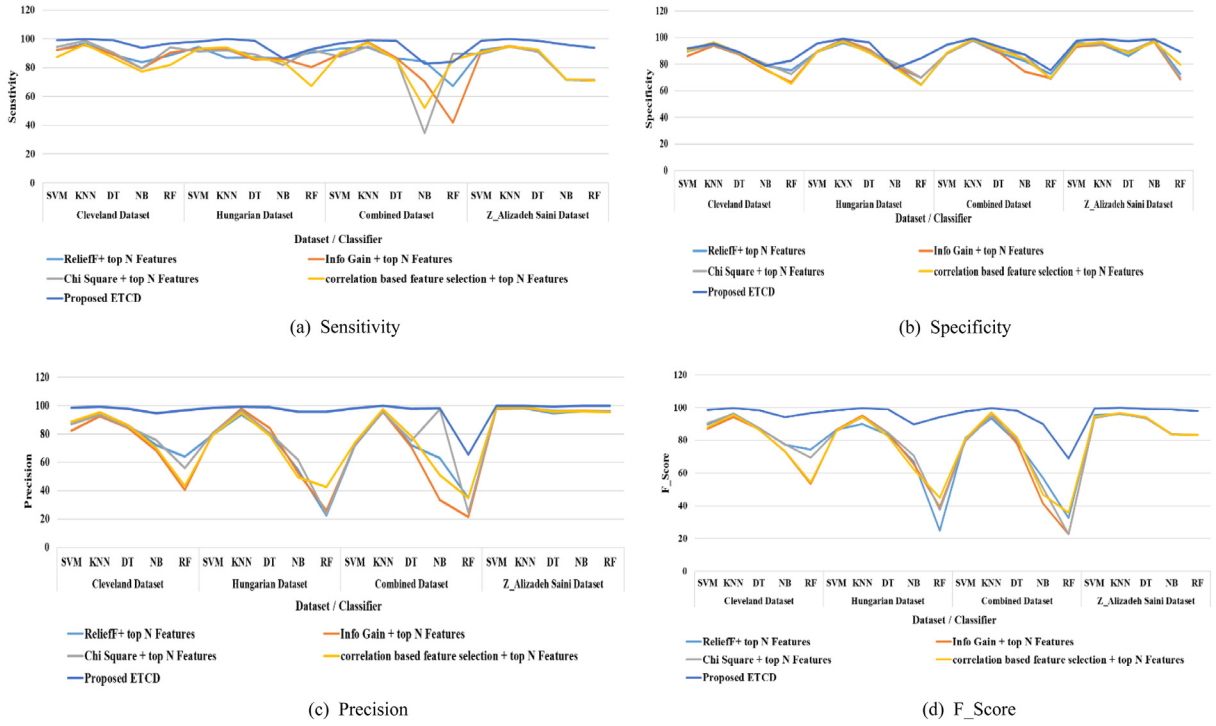


Fig. 8. Performance of ETCD in terms of (a) Sensitivity, (b) Specificity, (c) Precision, and (d) F_Score for all datasets with considered classifiers and FS Techniques.

and Hungarian datasets. Along with this ETCD gives >95% sensitivity in 75% of cases. On the other side, Fig. 8(b) represents specificity, which correctly identifies all patients who do not have the disease. With ETCD, specificity is greater than 80% with all classifiers and for all datasets. It is lowest with correlation-based FS and for RF classifiers. Figs. 8(c) and 8(d) show the precision and F_score performance of various methods and it is clear that ETCD increases the precision value by 19.59% for all datasets with all classifiers (Fig. 8(c).) Similarly, the increase in F_score with ETCD is 17.31% as shown in Fig. 8(d).

5.2.3.3. Performance comparison of ETCD with other FS techniques w.r.t. classifiers. Table 13 presents the comparison of classifier performance with respect to an average accuracy of FS techniques and ETCD accuracy for the above-mentioned datasets. For the Cleveland dataset, the performance of SVM, KNN, DT, NB, and RF gets increased by 7.011%, 2.64%, 6.875%, 11.459%, and 17.611% respectively. The performance of SVM, KNN, DT, NB, and RF gets increased by 5.359%, 2.263%, 8.121%, 1.734%, and 24.612% respectively for the Hungarian dataset. For the Combined dataset, the performance accuracy of SVM, KNN, DT, NB, and RF gets increased by 7.051%, 1.517%, 8.577%, 18.351%, and 8.065% respectively. Whereas, for the Z_Alizadeh Saini dataset, the performance accuracy of SVM, KNN, DT, NB, and RF gets increased by 5.486%, 4.460%, 5.809%, 24.794%, and 25.396% respectively. Overall, the ETCD improves SVM performance by 6.227%, KNN performance by 2.72%, and DT performance by 7.345%. On another side, NB and RF provide 14.084% and 18.921% more accuracy with ETCD. The OFSSA took more time to select features, which further depends upon the dimensionality of the dataset. But, the percentage of feature selection with OFSSA is very less as compared to other FS methods (Fig. 6). Further, Fig. 9 represents the average time taken by classifiers for various datasets with top N features, where N is 50%, 60%, and 70%., and features selected by OFSSA in proposed ETCD. After getting the appropriate features, the average time taken by classifiers in the proposed method is comparably less than others (Fig. 9).

5.2.3.4. Performance evaluation of ETCD with state of art methods. Additionally, the performance of the proposed approach ETCD in terms of accuracy was compared with other existing approaches for the above-mentioned datasets (Tables 14, 15 and 16). Table 14 compares the accuracy of the Cleveland and Hungarian datasets and it can be concluded that the proposed approach ETCD performs better with SVM, KNN, DT, and RF classifier for the Cleveland dataset, and for the Hungarian dataset, performance is better with SVM, KNN, and DT classifiers. Further, Table 15 compares the performance of the combined dataset. It is noted that, for the combined dataset, Ghosh et al. [11] give the best accuracy of 98.05%, 99.05%, and 98.32% with KNNBM, RFBM, and GBBM but with 10 features. Whereas ETCD gives an accuracy of 99.287% with 9 features. For the Z_Alizadeh Saini dataset, ETCD performs better with all classifiers as compared to existing results, given in Table 16.

5.3. Discussion

In this study, we proposed an efficient technique for cardiac disease prediction (ETCD) with optimal feature subset selection.

Table 13
Percentage difference in average accuracy with respect to Classifiers.

Dataset	Performance	SVM	KNN	DT	NB	RF
Cleveland dataset	Average accuracy	91.0075	96.6175	90.5825	79.9525	77.89
	ETCD Performance	97.87	99.24	97.27	90.3	94.54
	%age Difference	7.011852	2.642584	6.875193	11.45903	17.61159
Hungarian dataset	Average accuracy	92.265	96.935	90.225	81.885	68.625
	ETCD Performance	97.49	99.18	98.2	83.33	91.03
	%age Difference	5.359524	2.263561	8.121181	1.734069	24.61277
Combined dataset	Average accuracy	89.4725	97.675	88.9725	68.1275	70.9825
	ETCD Performance	96.26	99.18	97.32	83.44	77.21
	%age Difference	7.051215	1.517443	8.577374	18.35151	8.065665
Z_Alizadeh Saini dataset	Average accuracy	93.2475	95.215	92.41	72.115	71.5375
	ETCD Performance	98.66	99.66	98.11	95.89	95.89
	%age Difference	5.486013	4.460165	5.809805	24.79403	25.39629
Average %age difference		6.227151	2.720938	7.3458	14.08466	18.92158

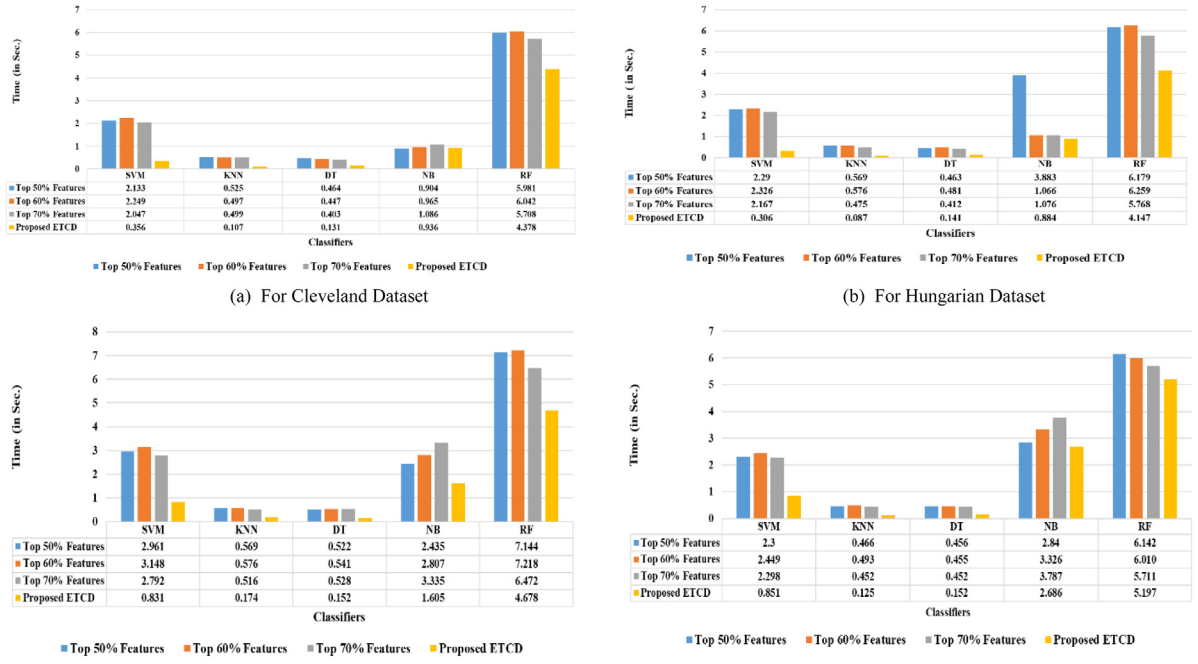
Table 14
Accuracy evaluation of ETCD with state-of-art methods for Cleveland and Hungarian datasets.

Author	Technique	Cleveland dataset	Hungarian dataset
Reddy et al. [32]	AGAFL	90%	91%
Gadekallu and Khare [33]	(CS+RS)+RS	91%	91.5%
Nourmohammadi-Khiarak et al. [7]	ICA+KNN	91.03% ±6.45%	–
Paul et al. [34]	Adaptive FDSS	92.31%	**95.56%**
Nasarian et al. [35]	2HFS	–	83.94%
Subramaniyam, Mahapatra, and Singh [36]	TGD + ACNN	92.52%	82.55%
Saqlain et al. [37]	FSSA	81.19%	84.582%
Arabasadi et al. [38]	GA+NN	89.4%	87.1%
El-Bialy et al. [39]	C4.5	78.54%	78.57%
Mokeddem [40]	FuzzyCDSS	90.5%	85.71%
Javeed et al. [41]	RSA-RF	**93.33%**	–
Li et al. [42]	FCMIM+SVM	92.37%	–
Gokulnath & Shantharajah [16]	GA+SVM	88.34%	–
Proposed ETCD +	SVM	**97.879%**	**97.491%**
	KNN	**99.242%**	**99.187%**
	DT	**97.273%**	**98.208%**
	NB	90.303%	83.333%
	RF	**94.545%**	91.039%

Table 15
Accuracy evaluation of ETCD with state-of-art methods for Combined datasets.

Author	Technique	Combined dataset
Mohan et al. [5]	HRFLM	88.7%
Dinesh et al. [26]	LR	86.51%
Ghosh et al [11]	KNNBM	98.05% (10 Features)
	RFBM	**99.05% (10 Features)**
	GBBM	98.32% (10 Features)
	DTBM	90.22% (10 Features)
Proposed ETCD +	SVM	97.268% (9 features)
	KNN	**99.287% (9 features)**
	DT	**98.321% (9 features)**
	NB	83.445% (9 features)
	RF	87.211% (9 features)

This proposed technique tries to enhance the efficacy of different classifiers for considered datasets with optimal feature subsets. For analysis, we considered four datasets of varying nature having varying dimensionality, four different feature selection methods ReliefF, Info-Gain, Chi-Square, and Correlation based feature selection, and five different classifiers SVM, KNN, DT, NB, and RF. First, we analyse the performance with data balancing. The result shows that balancing improves the performance of various datasets as well as classifiers performance (Table 5). Therefore, it is concluded that proper data balancing improves the performance of the model. After that, to analyse the impact of feature selection, we consider different feature ranking methods and compare the performance of all these methods based on the 'Top N Strategy'. It is concluded that, first, all features get different ranks with different techniques (Table 6). Second, it is very difficult to choose an optimal N for different feature selection methods and classifier performance varies with different feature subsets (Tables 7–10). Therefore, it is challenging to develop a method, which chooses the best optimal features with which all classifiers perform well. ETCD is able to deal with all these issues. It utilizes the optimal feature subset selection algorithm (OFSSA), which selects the best features from datasets, and with this chosen feature subset, all considered classifiers perform well (Table 11). A paired t-test with a significance level of 5% was also performed to statistically validate the ETCD results. The following two hypotheses have been put forth: H_o: "There is no major performance difference between the approaches" and H_a: "There is a substantial performance difference between the approaches". Table 12 depict that, the null hypothesis H_o can be rejected and the alternate hypothesis H_a can be accepted, as the paired t-test statistics are extremely significant with $p < 0.005$. Further, Table 13 presented the consistent improvement in the average accuracy of different classifiers for all datasets, giving a generalization capability to the predictive system. Also, from Tables 14–16, the proposed ETCD outperforms as compared to state-of-art methods.



(a) For Cleveland Dataset

(b) For Hungarian Dataset

(c) For Combined Dataset

(d) For Z_Alizadeh Saini Dataset

Fig. 9. Average time taken by classifiers after feature selection.

Table 16
Accuracy evaluation of ETCD with state-of-art methods for Z_Alizadeh Saini datasets.

Author	Technique	Z_Alizadeh Saini dataset
Alizadehsani et. al. [43]	SMO (1-1. 2-1, and 3-1)	92.41% (average)
Alizadehsani et. al. [12]	SMO	94.08%
Alizadehsani et. al. [44]	SVM	86.14% (LAD) , 83.17% (LCX), 83.50% (RCA)
Qin et al. [13]	EA-MFS	93.70%
Vijayashree & Sultana [14]	PSO	88.22%
Abdar et al. [45]	NE-nu-SVC	**94.66**%
Acharya et al. [46]	N2GC-nuSVM	93.08%
Nasarian et al. [35]	2HFS	92.58%
Proposed ETCD +	SVM	**98.668**%
	KNN	**99.666**%
	DT	**98.113**%
	NB	**95.893**%
	RF	**95.893**%

6. Conclusion and future scope

An efficient machine learning based technique for cardiac disease prediction called ETCD is described in the proposed work. The suggested framework ETCD can be used to anticipate instances in either healthy people or people with cardiac disease. ETCD uses an optimal feature subset selection approach (OFSSA) to choose optimal features from the cardiac disease datasets and train machine learning predictive models for instance categorization as cardiac disease and normal subject prediction. SVM, KNN, DT, NB, and RF classifiers are considered for classification. We use four datasets of varying nature from the UCI repository (Cleveland, Hungarian, Combined dataset (combination of four datasets), and Z Alizadeh Saini datasets) to validate the ETCD, and different feature ranking methods, ReliefF, Info gain, Chi-Square, and Correlation-based feature selection, to validate the performance of OFSSA. For feature ranking methods, we considered the "Top N Strategy" to select the features for classification. ETCD utilizing OFSSA performs well (Table 12) with less number of features (Fig. 6) as compared with other feature raking methods. ETCD improves the performance of classifiers in terms of used performance metric (Acc, Sens, Spec, Precision, F_Score) with the best features. With ETCD, the average performance for considered datasets with SVM, KNN, DT, NB, and RF classifiers get increased by 6.227%, 2.72%, 7.345%, 14.084%, and 18.921% respectively. We also performed a statistical paired t-test to compare the proposed ETCD's performance to other commonly utilized techniques. The results clearly demonstrate the ETCD's consistency. The result shows that it is possible to create a more accurate model for cardiac disease prediction by applying the proposed methodology and can be used by clinicians and healthcare professionals to detect heart disease in new patients, provided that patient data for the features used are available. In the future, the research will be expanded to include a multi-class classification of cardiac disease. We also intend to test our suggested approach with huge datasets with more features, as well as other kinds of features, such as ECG signals, etc. The authors also intend to apply the

proposed methodology to the diagnosis of other chronic diseases such as diabetes, cancer, and chronic kidney disease.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

(1) This material is the authors' own original work, which has not been previously published elsewhere.

(2) The paper is not currently being considered for publication elsewhere.

(3) The paper reflects the authors' own research and analysis in a truthful and complete manner.

(4) The paper properly credits the meaningful contributions of co-authors and co-researchers.

(5) All authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.

References

[1] P.A. Heidenreich, et al., Forecasting the future of cardiovascular disease in the United States: a policy statement from the American heart association, Circulation 123 (8) (2011) 933–944, http://dx.doi.org/10.1161/CIR.0b013e31820a55f5.

[2] A.L. Bui, T.B. Horwich, G.C. Fonarow, Epidemiology and risk profile of heart failure, Nat. Rev. Cardiol. 8 (1) (2011) 30–41, http://dx.doi.org/10.1038/nrcardio.2010.165.

[3] J. López-sendón, By J. López-Sendón, Spain, hear. Fail. Today a paradig, Shift 33 (4) (2011) 363–369.

[4] T. Tirkes, M.A. Hollar, M. Tann, M.D. Kohli, F. Akisik, K. Sandrasegaran, Response criteria in oncologic imaging: review of traditional and new criteria, Radiographics 33 (5) (2013) 1323–1341.

[5] S. Mohan, C. Thirumalai, G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques, IEEE Access 7 (2019) 81542–81554, http://dx.doi.org/10.1109/ACCESS.2019.2923707.

[6] B.A. Tama, S. Im, S. Lee, Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble, Biomed. Res. Int. 2020 (2020) http://dx.doi.org/10.1155/2020/9816142.

[7] J. Nourmohammadi-Khiarak, M.R. Feizi-Derakhshi, K. Behrouzi, S. Mazaheri, Y. Zamani-Harghalani, R.M. Tayebi, New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection, Health Technol. (Berl). 10 (3) (2020) 667–678, http://dx.doi.org/10.1007/s12553-019-00396-3.

[8] N.L. Fitriyani, M. Syafrudin, G. Alfian, J. Rhee, HDPM: An effective heart disease prediction model for a clinical decision support system, IEEE Access 8 (2020) 133034–133050, http://dx.doi.org/10.1109/ACCESS.2020.3010511.

[9] J. Nourmohammadi-Khiarak, M.-R. Feizi-Derakhshi, F. Razeghi, S. Mazaheri, Y. Zamani-Harghalani, R. Moosavi-Tayebi, New hybrid method for feature selection and classification using meta-heuristic algorithm in credit risk assessment, Iran J. Comput. Sci. 3 (1) (2020) 1–11, http://dx.doi.org/10.1007/s42044-019-00038-x.

[10] A. Gupta, R. Kumar, H. Singh Arora, B. Raman, MIFH: A machine intelligence framework for heart disease diagnosis, IEEE Access 8 (MI) (2020) 14659–14674, http://dx.doi.org/10.1109/ACCESS.2019.2962755.

[11] P. Ghosh, S. Azam, M. Jonkman, S. Shultana, A.R. Beeravolu, Efficient Prediction of Cardiovascular Disease using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques, vol. 9, 2021, http://dx.doi.org/10.1109/ACCESS.2021.3053759.

[12] R. Alizadehsani, J. Habibi, M. Javad, B. Bahadorian, Z. Alizadeh, A data mining approach for diagnosis of coronary, Comput. Methods Programs Biomed. (2013) 1–10, http://dx.doi.org/10.1016/j.cmpb.2013.03.004.

[13] C. Qin, Q. Guan, X. Wang, Application of Ensemble Algorithm Integrating Multiple Criteria Feature Selection in Coronary Heart, Vol. 29, (6) 2017, pp. 1–11, http://dx.doi.org/10.4015/S1016237217500430.

[14] J. Vijayashree, H.P. Sultana, A machine learning framework for feature selection in heart disease classification using improved particle swarm optimization with support vector machine classifier, Program. Comput. Softw. 44 (6) (2018) 388–397, http://dx.doi.org/10.1134/S0361768818060129.

[15] M.S. Amin, Y.K. Chiam, K.D. Varathan, Identification of significant features and data mining techniques in predicting heart disease, Telemat. Inform. 36 (2019) 82–93, http://dx.doi.org/10.1016/j.tele.2018.11.007.

[16] C.B. Gokulnath, S.P. Shantharajah, An optimized feature selection based on genetic approach and support vector machine for heart disease, Cluster Comput. 22 (2019) 14777–14787, http://dx.doi.org/10.1007/s10586-018-2416-4.

[17] S. Wadhawan, R. Maini, A systematic review on prediction techniques for cardiac disease, Int. J. Inf. Technol. Syst. Approach 15 (1) (2022) 1–33.

[18] J. Mishra, S. Tarar, Chronic Disease Prediction using Deep Learning BT - Advances in Computing and Data Sciences, 2020, pp. 201–211.

[19] M. Tarawneh, O. Embarak, Hybrid Approach for Heart Disease Prediction using Data Mining Techniques BT - Advances in Internet, Data and Web Technologies, 2019, pp. 447–454.

[20] I. Kadi, A. Idri, J.L. Fernandez-Aleman, Knowledge discovery in cardiology: A systematic literature review, Int. J. Med. Inform. 97 (2017) 12–32, http://dx.doi.org/10.1016/j.ijmedinf.2016.09.005.

[21] P. Sharma, K. Choudhary, K. Gupta, R. Chawla, D. Gupta, A. Sharma, Artificial plant optimization algorithm to detect heart rate & presence of heart disease using machine learning, Artif. Intell. Med. 102 (2020) http://dx.doi.org/10.1016/j.artmed.2019.101752.

[22] I. Javid, A.K.Z. Alsaedi, R. Ghazali, Enhanced accuracy of heart disease prediction using machine learning and recurrent neural networks ensemble majority voting method, Int. J. Adv. Comput. Sci. Appl. 11 (3) (2020) 540–551, http://dx.doi.org/10.14569/ijacsa.2020.0110369.

[23] E. Nasarian, et al., Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach, Pattern Recognit. Lett. 133 (2020) 33–40, http://dx.doi.org/10.1016/j.patrec.2020.02.010.

[24] N. Kumar, K. Sikamani, Prediction of chronic and infectious diseases using machine learning classifiers—A systematic approach, Int. J. Intell. Eng. Syst. 13 (4) (2020) 11–20.

[25] C.B.C. Latha, S.C. Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Inform. Med. Unlocked 16 (2018) 100203, http://dx.doi.org/10.1016/j.imu.2019.100203, 2019.

[26] K.G. Dinesh, K. Arumugaraj, K.D. Santhosh, V. Mareeswari, Prediction of cardiovascular disease using machine learning algorithms, in: 2018 International Conference on Current Trends Towards Converging Technologies, ICCTCT, 2018, pp. 1–7, http://dx.doi.org/10.1109/ICCTCT.2018.8550857.

[27] Purushottam K. Saxena, R. Sharma, Efficient heart disease prediction system, Procedia Comput. Sci. 85 (2016) 962–969, http://dx.doi.org/10.1016/j.procs.2016.05.288.

[28] W. Shah, et al., A machine-learning-based system for prediction of cardiovascular and chronic respiratory diseases, J. Healthc. Eng. 2021 (2021) http://dx.doi.org/10.1155/2021/2621655.

[29] F.I. Alarsan, M. Younes, Analysis and classification of heart diseases using heartbeat features and machine learning algorithms, J. Big Data 6 (1) (2019) http://dx.doi.org/10.1186/s40537-019-0244-x.

[30] S.I. Sherly, G. Mathivanan, An ensemble basedheart disease predictionusing gradient boosting decision tree, Turkish J. Comput. Math. Educ. 12 (10) (2021) 3648–3660, [Online]. Available: https://www.proquest.com/scholarly-journals/ensemble-basedheart-disease-predictionusing/docview/2628341266/se-2.

[31] S. Wadhawan, R. Maini, EBPSO: Enhanced binary particle swarm optimization for cardiac disease classification with feature selection, Expert Syst. (2022) e13002.

[32] G.T. Reddy, M.P.K. Reddy, K. Lakshmanna, D.S. Rajput, R. Kaluri, G. Srivastava, Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis, Evol. Intell. 13 (2) (2020) 185–196, http://dx.doi.org/10.1007/s12065-019-00327-1.

[33] T.R. Gadekallu, N. Khare, Cuckoo search optimized reduction and fuzzy logic classifier for heart disease and diabetes prediction, Int. J. Fuzzy Syst. Appl. 6 (2) (2017) 25–42, http://dx.doi.org/10.4018/IJFSA.2017040102.

[34] A.K. Paul, P.C. Shill, M.R.I. Rabin, K. Murase, Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease, Appl. Intell. 48 (7) (2018) 1739–1756, http://dx.doi.org/10.1007/s10489-017-1037-6.

[35] E. Nasarian, et al., Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach, Pattern Recognit. Lett. 133 (2020) 33–40, http://dx.doi.org/10.1016/j.patrec.2020.02.010.

[36] A. Subramaniyam, R.P. Mahapatra, P. Singh, Taylor and gradient descent-based actor critic neural network for the classification of privacy preserved medical data, Big Data 7 (3) (2019) 176–191, http://dx.doi.org/10.1089/big.2018.0166.

[37] S.M. Saqlain, et al., Fisher score and matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines, Knowl. Inf. Syst. 58 (1) (2019) 139–167, http://dx.doi.org/10.1007/s10115-018-1185-y.

[38] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, A.A. Yarifard, Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm, Comput. Methods Programs Biomed. 141 (2017) 19–26, http://dx.doi.org/10.1016/j.cmpb.2017.01.004.

[39] R. El-Bialy, M.A. Salamay, O.H. Karam, M.E. Khalifa, Feature analysis of coronary artery heart disease data sets, Procedia Comput. Sci. 65 (Iccmit) (2015) 459–468, http://dx.doi.org/10.1016/j.procs.2015.09.132.

[40] S.A. Mokeddem, A fuzzy classification model for myocardial infarction risk assessment, Appl. Intell. 48 (5) (2018) 1233–1250, http://dx.doi.org/10.1007/s10489-017-1102-1.

[41] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, R. Nour, An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection, IEEE Access 7 (2019) 180235–180243, http://dx.doi.org/10.1109/ACCESS.2019.2952107.

[42] 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 8 (Ml) (2020) 107562–107582, http://dx.doi.org/10.1109/ACCESS.2020.3001149.

[43] R. Alizadehsani, M.J. Hosseini, Z.A. Sani, A. Ghandeharioun, R. Boghrati, Diagnosis of coronary artery disease using cost-sensitive algorithms, in: 2012 IEEE 12th International Conference on Data Mining Workshops, 2012, pp. 9–16.

[44] R. Alizadehsani, et al., Coronary artery disease detection using computational intelligence methods, Knowl. Based Syst. 109 (2016) 187–197.

[45] M. Abdar, U.R. Acharya, N. Sarrafzadegan, V. Makarenkov, NE-nu-SVC: A new nested ensemble clinical decision support system for effective diagnosis of coronary artery disease, IEEE Access (2019) 1, http://dx.doi.org/10.1109/ACCESS.2019.2953920.

[46] U.R. Acharya, R. Tan, V. Makarenkov, P. Pławiak, Computer Methods and Programs in Biomedicine a New Machine Learning Technique for an Accurate Diagnosis of Coronary Artery Disease, Vol. 179, 2019, http://dx.doi.org/10.1016/j.cmpb.2019.104992.