Detection of Cardio Vascular Disease Using Ensemble Feature Engineering with Decision Tree

Debasmita Ghosh Roy[1], João Manuel R. S. Tavares[2]
[1]*School of Automation, Banasthali Vidyapith, Rajasthan- 304022, India*

[2]*Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal*

ABSTRACT

Cardiovascular diseases (CVDs) are a cluster of heart-related issues, including many comorbidities, which is becoming a leading cause of human death across the globe. Hence, an essential framework is demanded for the early detection of CVDs which can help to prevent premature death. The application of Artificial Intelligence (AI) in healthcare has opted for this challenge and makes it easier to detect CVDs using a computational model. In this study, authors built a reduced dataset using ensemble feature selection methods and got five features as per their weight values. Support Vector Machine, Logistic Regression, and Decision Tree classification techniques are used to check the effectiveness of newly designed datasets through different validation approaches. The authors also worked on data processing and visualization techniques, including Principal Component Analysis (PCA), and T-sne for understanding the data structure. From the findings, it was possible to conclude that DT has achieved an optimal accuracy and AUC of 98.9% and 0.99 ROC with leave one out Cross Validation (CV).

Keywords: Ensemble feature engineering, Classification, AI, CVDs

INTRODUCTION

Cardiovascular disease (CVDs) is one of the deadliest diseases in the world. In 2019, WHO and GBD reported that 30 % deaths occurred due to CVDs and it is also expected that 23.6 million people may get affected by the end of 2030 (Joshi & Rienks 2021). In the last few decades, the number of death due to CVDs has increased noticeably from 15.2% to 28.1% in India (Tichenor and Sridhar 2019). Indian public medical system is still incapable of preventing non-communicable diseases including CVDs. Due to lack of the five A's like awareness, access, absence, affordability, and accountability in India, the efficient healthcare facility is not adequate (Kasthuri et al 2018). Medical investigation and therapeutic assistance are still unavailable in rural areas which hamper human health. Despite the government initiatives of medicare for needy people, the majority of Indian populace does not have preventive health care benefits.

All these factors are associated with slow therapy and increased morbidity and mortality (Peiris & Ghosh 2021). Similarly, few developed countries including China, America has a higher percentage of CVDs death. As a result, CVDs are a prime reason for health loss for all regions of the world (GRF Collaborators 2018, Roth & Mensah 2020).

Cardiovascular diseases are a cluster of disorders of heart and blood vessels, which are typically associated with human behavioural factors (Jennum et al 2021). The most significant behavioural risk factors include poor diet, physical latency, tobacco and liquor consumption. The effect of hazardous factors may increase blood pressure, glucose level, blood lipids, overweight and obesity. These factors indicate an increase in cardiac disorder and other complexity (Sabzmakan & Morowatisharifabad 2014). The most accurate method for CVDs diagnosis is coronary computed tomography angiogram, echocardiogram, and magnetic resonance imaging. These tests are relatively expensive, time-consuming and not readily available in low resource regions (Abbara & Blanke 2016). To address these issues, an intelligence model is highly demanded which is used for accurate and early detection of CVDs. Therefore, the availability of such services is essential to save physicians time and reduce the financial burden for unprivileged patients. Additionally, this model acts as a pre-alarming system for normal people who have a probability of growing cardiac disorder (Haq et al 2021).

Artificial intelligence played an important role to design an intelligence model for disease diagnosis and prognosis. There are various factors that conduce to disease prognosis including cardiovascular disease, cancer, alzheimer and others (Davenport & Kalakota 2019, Dey & Ashour 2018, Louis & Gerber 2014). In the last few decades, scientists were more concentrated in recognizing the important features which need to be utilized in their CVDs prediction models (Singh & Sinha 2016, Kausar & Abdullah 2016). For this purpose, a systematic analysis like data pre-processing and feature engineering is required, and it is considered an essential task. Using these two folds, the key features and the significance of every feature with the target class can easily be found which helps to eliminate the unwanted features (Lysaght & Lim 2019, Dey & Ashour 2017). Machine learning effectively predicted the risk of heart diseases and many related studies were previously conducted. Previously, maximum scientists inquired to find the best classifier that gives fine CVDs prediction accuracy. Although, few researchers inquired to identify significant features associated with heart disease and their relationship (Suganya et al 2017). To the best of our learning, this is the first study where the strength score of each feature is used as a significant predictor for CVDs prediction.

In this study, our main objective was to choose the foremost and complimenting attributes. These characteristics will help the underlying classifier to produce a complex decision surface for distinguishing between the presence and absence of heart diseases. Our goal with the optimal feature set implemented for our CVDs classification scenario was of three-fold: *i)* improve the model efficacy of underlying classifiers, *ii)* Create an optimal feature set to explain the CVDs association, *iii)* Comparison analysis of our results with the ones previously reported in related studies where the same dataset was used. In addition to the primary objective to figure out the optimal feature set providing elevated classification efficacy, our second aim was to choose a quick and well-performing classifier such as a random forest (Singh & Sinha 2016, Rajdhan & Agarwal 2020). The implementation of feature engineering will also make it possible to an overall lesser time, for this reason only a lessened number of attributes is to be extracted and utilized in the classification process.

In this article, we proposed a computerized decision support system that is capable to detect the presence and absence of CVDs from clinical data analysis using artificial intelligence. The remainder of the article is structured as follows: recent research articles are discussed in Section II. Section III contains the methods including dataset description, feature selection, and classification. Experimental outcomes involving the evaluation protocol and the obtained results are discussed in Section IV. Finally, a summary highlighting

the main findings of our study is provided in Section V, which are followed by limitations and future direction in Section VI.

BACKGROUND

Salhi et al. (Salhi et al. 2021) used three classification techniques namely NN, SVM, and K-NN and the optimal accuracy of 93% was achieved by NN. Gao et al. (Gao et al. 2021) implemented an ensemble model and compared the result with six different classifiers. The bagging ensemble learning method with DT outperforms and obtained accuracy was 98.6%. The authors also used two feature extraction methods: LDA and PCA to select the significant features among all features. Gárate-Escamila et al. (Gárate-Escamila et al. 2020) utilized six classifiers namely DT, GBT, LOG, MPC, NB, and RF for classification of CVD. The authors used Chi-square and PCA to select the important features among 74 features. The maximum efficacy was achieved by chiSqSelector +PCA and RF. Spencer et al. (Spencer et al. 2020) used four feature selection methods such as chi-square, relief, symmetrical uncertainty, and PCA to create a distinctive feature set. Eight classification algorithms are used to verify the optimal feature set contribution to heart disease prediction. Finally, the maximum efficacy is achieved by chi-square feature selection with bayesnet algorithm. Almustafa et al. (Almustafa et al. 2020) used seven classification algorithms and the maximum efficacy was achieved by K-NN (n=1). Bashir et al. (Bashir et al. 2019) used five machine learning techniques: DT, LR, LR (SVM), NB, and RF to improve the classification accuracy of heart disease prediction. Of all, an optimal accuracy of 84.85% was achieved by SVM. Khourdifi et al. (Khourdifi et al. 2019) used five algorithms such as K-NN, SVM, RF, NB, and MLP followed by two optimization process (FCBF, FCBF+ACO+PSO). The optimal accuracy was achieved by KNN with FCBF+ACO+PSO. Alotaibi et.al (Alotaibi et.al 2019) used five algorithms, namely NB, DT, RF, LR and SVM to predict heart failure disease. Of all, maximum accuracy was achieved by DT. Haq et al. (Haq et al. 2018) implemented a hybrid intelligence system to predict heart disease and compared the performance with six classifiers. The authors used three feature selection methods: minimal-redundancy-maximal-relevance, least absolute shrinkage and selection operator, relief to identify the significant feature. The optimal efficacy was obtained by LR with relief feature predictor. Kodati et al. (Kodati et al. 2018) used four classification algorithms: NB, SVM, K-NN, and RF. The authors reported that NB classifier worked well as compared to the other classifiers. Hung et al. (Hung et al. 2018) used three classification algorithms: SVM (kernel type: linear, gaussian, polynomial, sigmoid), NB, and LR to diagnose the heart disease. The maximum efficacy was achieved by SVM with linear kernel type. The authors also used three feature selection methods PCA, CFS, LLCFS, ILFS to identify significant features. Finally, ILFS is the best feature selection technique as compared to others. Uyar and İlhan (Uyar & Ilhan 2017) implemented a genetic algorithm based on trained the recurrent fuzzy neural network to diagnose heart disease and the obtained efficacy is 97.78%. Kumar et al. (Kumar et al. 2017) used three feature selection method: relief, recursive, boruta for identifying each feature importance to contribute heart disease diagnosis. The author used RF algorithm where the optimal accuracy is 98% and boruta confirmed features. Pahwa & Kumar (Pahwa & Kumar 2017) used a hybrid feature selection method (SVM+RFE) to create a feature subset. Two classification techniques like RF, NB are used to check the effectiveness of the new feature subset. Finally, the accuracy obtained by NB and RF classifier is 84.15% and 84.16% respectively. Suganya & Rajaram (Suganya & Rajaram 2017) applied two machine learning algorithms such as SVM, NN for the prediction of heart disease. The author also applied a hybrid feature selection method and the reported optimal accuracy of 97% was achieved by SVM. Jabbar & Deekshatulu (Jabbar & Deekshatulu 2016) proposed a hybrid model by using RF classifier along with chi square feature section. Four classification algorithms such as PART C4.5, NB, DT, and NN were used for comparison with the proposed model. The optimal accuracy of 83.70% was achieved by the proposed model. Dey & Singh (Dey & Singh 2016) used PCA to find the minimum number of features and three classification algorithms such as DT, SVM, NB to predict heart disease. The maximum accuracy was achieved by SVM. Feshki and Shijani (Feshki and Shijani 2016) used PSO algorithm to find the most important features for heart disease. Four algorithms such as SVM, C4.5, MLP, FFBF are used and 91.94% accuracy was achieved by FFBF. Khan et al. (Khan et al. 2016) used five classification

techniques namely zeroR, oneR, J48, RF, and MLP to predict heart disease. Of all, MLP provided good performance over the others with kappa statistics of 0.6705. Khanna et al. (Khanna et al. 2015) used three classification methods: LR, SVM, and NN. The maximum reported accuracy of 88.2% was obtained by LR. The summarization of earlier work related to CVDs prediction is represented in Table 1.

Table 1. Background of study

Authors (year)	Classifiers	Compared with	Performance (%)			
			Acc	Sen	Spec	AUC
Salhi et al. (2021)	NN	SVM, KNN	93.0	-	-	-
Gao et al. (2021)	Bagging ensemble learning method with decision tree	KNN, SVM, NB, DT, RF	98.6	-	-	-
Gárate-Escamila et al. (2020)	chiSqSelector +PCA and RF	DT, GBT, LOG, MPC, NB, and RF	98.7	-	-	-
Spencer et al. (2020)	BayesNet	LR, SGD, IBK, AdaB (DS), AdaB(Logistic), JRip,RF	85	-	-	-
Almustafa et al. (2020)	K-NN	NB, SGD, SVM, DT, Adaboost, JRip, J48	99.7073	-	-	-
Bashir et al. (2019)	LR(SVM)	DT, LR, NB, RF	84.85	-	-	-
Khourdifi et al. (2019)	K-NN +FCBF+ACO+PSO	SVM, RF, NB, and MLP	99.65	-	-	-
Alotaibi et.al (2019)	DT	LR, RF, NB, SVM	93.19	-	-	-
Haq et al. (2018)	LR+relief	K-NN, ANN, SVM, NB, DT, RF	89	-	-	-
Kodati et al. (2018)	NB	SVM, RF, K-NN	83.70	-	-	-
Hung et al. (2018)	SVM	LR, NB	90.65	87	93	96
Uyar & Ilhan (2017)	RFNN	-	97.78	-	-	-
Kumar et al. (2017)	RF	-	98	-	-	-
Pahwa & Kumar (2017)	RF	NB	84.16	-	-	-
Suganya & Rajaram (2017)	SVM	NN	97	-	-	-
Jabbar & Deekshatulu (2016)	RF+chi square	PART C4.5, NB, DT, NN	83.70	85.8	82.3	-
Dey & Singh (2016)	SVM	DT, NB	-	-	-	-
Feshki and Shijani (2016)	FFBF	SVM, c4.5, MLP	91.4	-	-	-

Khan et al. (2016)	MLP	zeroR, oneR, J48, RF	-	-	-	-
Khanna et al. (2015)	LR	SVM, NN	88.2			

METHODS

In this section, the distinct processing steps of the proposed model, beginning with feature description, selection, and classifications, which aids to identify the chances of CVDs, are explained.

Feature Description

Table 2. Description of the used features

No.	Attribute	Description
1	Age	Age (in years)
2	Gender	1 - Male 0 - Female
3	Cp	Chest pain type: 1 - Typical Angina 2 - Atypical Angina 3 - Non-Angina Pain 4 - Asymptomatic Pain
4	Trestbps	Resting blood pressure (in mm Hg)
5	Chol	Serum cholesterol (in mm/dl)
6	Fbs	Fasting blood sugar > 120 mg/dl 1 - true 0 - false
7	Rest ECG	Resting electrocardiographic results Value 0 - Normal Value 1 - Having ST-T wave abnormality Value 2 - Showing probable or define left ventricular hypertrophy
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina 1 - yes 0 - no
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	The slope of the peak exercise ST segment: 1 - Up sloping 2 - Flat 3 - Down sloping
12	Ca	Number of major vessels (0-3) colored by fluoroscopy
13	Thal	Thalassemia value: 3 - normal

		6 - fixed defect 7 - reversable defect
14	Target	Patient have heart disease=1 Patient not have heart disease =0

We considered the 4 databases such as Cleveland, Hungary, Switzerland, and VA long beach which was obtained from the University of California Irvine (UCI) repository to assess the performance of the proposed model. The original dataset consists of 303 instances and 75 attributes where the target class is assigned from 0 (zero) to 1-4. However, most of the researchers have chosen 14 attributes, which is a subset of 14 features (UCI). Three types of features were present: *i)* categorical, *ii)* integer and *iii)* real. The clinical variables considered relevant were AGE, SEX, CP, and TRESTBPS; the routine test data CHOL, FBS, and RESTECG; the exercise electrocardiography test with the features THALACH, EXANG, SLOPE, and OLD PEAK; and the non-invasive test, THAL, and CA, Table 2. Among 303 instances, 165 and 138 instances belong to 0 (absence of CVD) and 1-4 (presence of CVD), respectively. Hence, the dataset was imbalanced, Figure. 1 and Table 3. In this study, we used the target class as 0 (zero) and 1 (one), which are associated with binary classification tasks. In addition, we used PCA and T-sne to visualize the data in two dimensions, Figure. 2.

Table 3. Dataset distribution

Dataset Name	Total number of instances	Presence of CVDs	Absence of CVDs
Cleveland, Hungary, Switzerland, and VA long beach	303	138 (45%)	165 (55%)

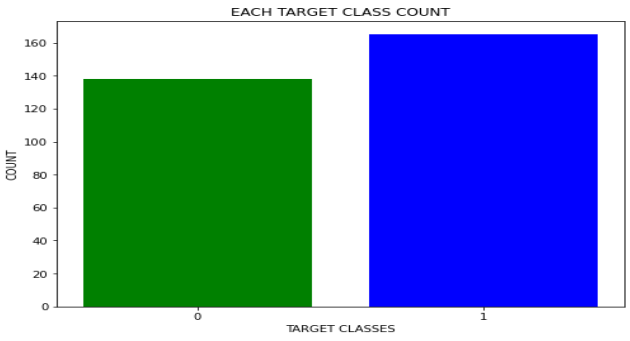


Figure 1. Structure of the used dataset

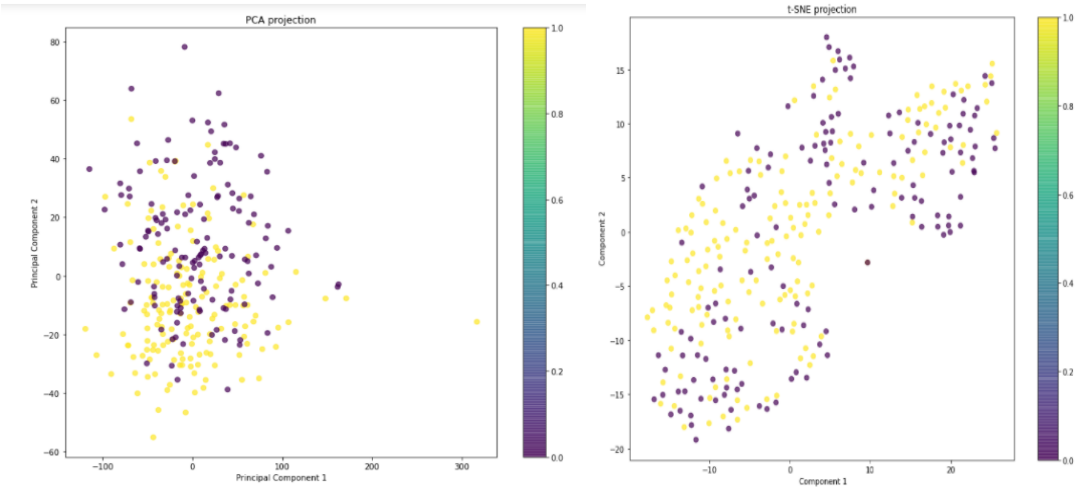


Figure 2. Visualization in 2d of the used dataset

Feature Selection and Classification

Feature selection (Janecek & Gansterer 2008) reduces the number of input features when developing a clinical decision support system (CDSS). In the literature, many systems are devoted for CVD analysis (Baranidharan & Pal 2019, Kausar & Palaniappan 2016) where the authors do not specify the motive for their selection for specific features. Instead, they are chosen based on well-known feature reduction techniques like PCA, Chi, and linear discriminant analysis (LDA). However, the selection of features is necessary and helpful in designing efficient classification models. For this purpose, we considered ensemble (Archer & Kimes 2008), embedded feature selection (Menze & Kelm 2009) along with conventional technique (Gao & Ali 2021). Also, instead of using single reduction or selection techniques, we conducted the selection process based on weightage value, which is achieved by individual techniques. Due to the utilization of different methods, the corresponding features score were varied in order to establish an optimal feature set. The features subset providing better accuracy was retained as the winner combination.

For classification, DT, SVM, LR classifiers were considered. Our choice for the selection of these classifiers was not arbitrary. We conducted a literature survey for disease diagnosis and based on our problem, SVM, DT, and LR were identified as well performed classifiers where the target label is binary (Mythili & Mukherji 2013, Pandey & Pandey 2013). Each has individual benefits, SVM, and LR provides a clear surface line between disease or non-disease class labels whereas in case of DT, the choices are set out logically, which is easy to understand. However, before selecting the classifier, we performed a preliminary experiment using SVM, DT, LR, and RF on the actual dataset with eight validation techniques. The result was quite surprising, RF classifier provided the best accuracy in all validation like holdout, 5-fold, 10-fold, and others. SVM, LR, and DT provided less accuracy compared to RF. To check the effectiveness of our feature set, we finally selected the weak classifiers in this study. The procedure followed in doing the same is shown in Figure. 4.

Evaluation Protocol

Several measures were considered for proper evaluation of the performance of the proposed model. Accuracy (ACC), sensitivity (SEN), specificity (SPEC), and area under the curve (AUC) were selected to

measure its performance. We also concentrated on the F1-score, which provides the harmonic average between precision and recall:

$$\text{F1 score: } 2\frac{(Precision \times Recall)}{(Precision+Recall)} \quad (1)$$

Similarly, AUC was necessary to understand the behaviour of underlying classifiers. The classification accuracy, sensitivity, and specificity are computed as:

$$\text{Accuracy: } \frac{(TP+TN)}{(TP+FN+FP+TN)} \quad (2)$$

$$\text{Sensitivity: } \frac{TP}{TP+FN} \quad (3)$$

$$\text{Specificity: } \frac{TN}{TN+FP} \quad (4)$$

where *TP* is the true positives, *TN* the true negatives, *FN* the false negatives, *FP* the false positives, and *TN* the true negatives.

Experimental Results and Discussion

This section details and discusses the experimental procedure, like validation techniques, including classification performance, feature importance, and classification performance with an optimal feature set. Finally, the outcomes obtained from different experimental setups are described followed by results comparisons.

Data Insights

Mainly three datasets are available; Cleveland dataset is one of the widely used dataset for heart disease diagnosis. The raw dataset comprises of 75 features and 303 instances which is divided into 7 groups: 1) patient record data, 2) medical during the exercise test, 3) exercise electrocardiogram, 4) cardiac fluoroscopy, 5) exercise thallium scintigraphy, 6) coronary angiograms, and 7) blood Vessels. Most of the authors have built their prediction model for heart disease prediction by 14 features with binary or multi-class target labels.

Classifier Selection

Several works were carried out for the detection of CVD using different ML classifiers. However, every classifier performed well, but most of the researchers concluded that random forest (Khalilia & Chakroborty 2011, Palimkar & Shaw 2022) decision tree (Rajdhan & Agarwal 2020, Singh & Kumar 2020) and support vector machine (Singh & Kumar 2020, Fitriyani & Syafrudin 2020) provide the best outcome as compared to others. For this reason, we selected these three classifiers and utilized them upon the raw dataset. In the next stage, after choosing the classifier, we have selected eight different validation techniques such as 5-fold, 10-fold, 20-fold, stratified 5-fold, 10-fold, 20-fold, leave one out, and hold out. Of all, random forest classifiers outperformed with an optimal accuracy except leave one out techniques. Similarly, SVM, DT, and LR provided the best accuracy with leave one out and hold out validation techniques. Our goal was to improve the efficacy of the weakest classifiers (SVM, DT, LR). In this study, the lowest efficacy was achieved by the decision tree along with seven validation techniques. So, we have selected DT classifier along with two validations such as leave one out and holdout approaches and move forward to our next

level of experiment. The selected classifier's performance, along with eight validation protocols is summarized in Table 4.

Features Importance

It is the most predominant factor for designing classification models. There are several types of measurement approaches that are utilized to identify the importance of each feature. The main goal of any classification model is to map the relationship between input and output factors. In this study, we referred to three feature selection methods namely extra tree classifier, PCA, and random forest classifier. These methods are implemented on raw datasets to explain the association between input features and output class labels individually.

Many pieces of literature exist where three methods are implemented individually for the selection of important feature associations for heart disease diagnosis. Among them, one can distinguish filter, ensemble, and embedded feature selection methods. Filter technique selects the features which are independent of any classifier. Ensemble methods are robust and stable, which reduces the biases and searches the optimal feature subset which helps to build a proper model. Although we are not focused on a single method, primarily we have compared each score value that is obtained by each selection method. In the next level of an experiment, we considered the weightage score value and figured out the most relevant features. In this manner, five features are selected from the raw dataset for creating a new feature subset. The designation of the five features is: Cp, Thalach, Oldpeak, Ca, and Thal and the score values are: 0.426638, 0.282536, 0.298668, 0.459829, and 0.371493 respectively. Individual feature importance values are listed in Table 5.

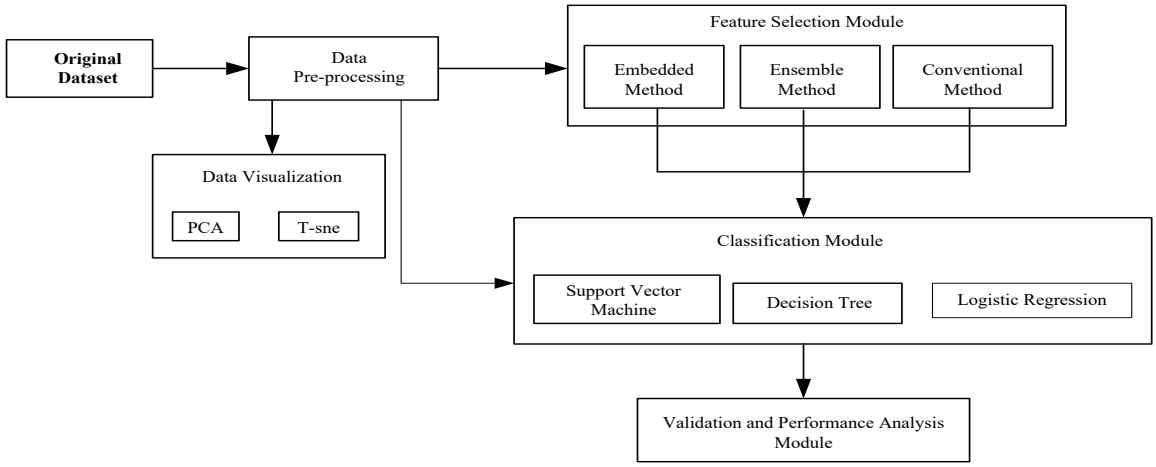


Figure 3. Flow chart of the proposed model for CVDs diagnosis

Table 4: Accuracy of the studied classifiers using different CVs (best values in bold)

No	Cross Validation	ACC (%)			
		SVM	DT	RF	LR
1	5-fold	80.63	75.21	83.92	79.37
2	10-fold	79.78	77.73	82.70	77.96

3	20-fold	81.02	79.90	83.46	81.42
4	Stratified 5-fold	80.62	73.95	83.80	78.61
5	Stratified 10-fold	80.62	73.95	83.47	78.72
6	Stratified 20-fold	79.78	79.38	82.47	80.04
7	Leave one out	83.83	78.55	82.82	90.12
8	Hold out	81.97	81.96	85.25	89.68

Table 5: Scores of features using different feature selection methods (best values in bold)

Features	PCA scores	RF scores	Extra Tree scores	Total scores
Age	0.03698	0.08787	0.07013031	0.19498
Sex	0.05151	0.0482	0.06937022	0.16908
Cp	0.19581	0.1198	0.11102771	**0.426638**
Trestbps	0.01426	0.07551	0.06498572	0.154756
Chol	0.01053	0.07767	0.06110335	0.149303
Fbs	0.00037	0.0102	0.0201098	0.03068
Restecg	0.00086	0.01997	0.03141138	0.052241
Thalach	0.0807	0.11308	0.08875593	**0.282536**
Exang	0.11519	0.06094	0.10081089	0.276941
Oldpeak	0.1049	0.10688	0.0868883	**0.298668**
Slop	0.03836	0.03915	0.05536537	0.132875
Ca	0.17901	0.13856	0.14225851	**0.459829**
Thal	0.17152	0.10219	0.09778251	**0.371493**

Model Design and Performance Estimation

We have trained the classification model after the selection of classifier and making a new feature subset. To estimate the performance of this model, the leave-one-out cross validation was applied. It is a special type of validation technique where the number of folds is equal to the number of instances in the dataset. In this way, the learning algorithm is implemented once for each instance, using all other instances as a training set and using the selected instances as a single-item test set. Additionally, it has no variance issue, so it is a better validation technique as compared to others.

Table 6: Comparing the accuracy achieved by the studied classifiers (best values in bold)

No of features	SVM					DT					LR				
	Performance (%)					Performance (%)					Performance (%)				
	Acc	Sen	Spe	F1 score	Auc	Acc	Sen	Spe	F1 score	Auc	Acc	Sen	Spe	Auc	F1score
13	83.83	84.6	81.6	84	81	78.55	83.4	81.6	83	82.29	90.1	91.7	89.7	87.01	87
5	90.89	90.7	87.4	89	83	**98.9**	96.5	98.1	91	99	94.8	94.3	92.07	93.1	90

To confirm the significance of feature set, the model performance was measured using accuracy and area under the curve value. We have compared the classification model performance based on number of features. Finally, the maximum accuracy (98.9%) was obtained by the decision tree classifier along with the new feature subset. The recorded accuracy of each classifier is listed in Table 6. We have also referred to area under the curve value due to an imbalanced dataset. The AUC value of 0.99 was achieved by DT.

Comparison of Existing CVDs Model with Cleveland Dataset

A benchmark was required to compare the performance of the existing models of CVD with the proposed intelligent model. Table 7 indicates the results of a comparison study between the existing models and our proposed model in terms of classification performance.

Table 7: Comparison with existing models in terms of classification performance

Authors (year)	Method	Features	Accuracy (%)
Alotaibi et al. (2019)	SVM	13	92.30
Pahwa et al. (2017)	SVM recursive feature elimination algorithm with NB	10	85.15
	SVM recursive feature elimination algorithm with RF	12	84.15
Suganya et al. (2017)	SVM Sequential minimal optimization	5	97
Khan et al. (2016)	RF	13	89.25
	DT	13	89.10
Khanna et al. (2015)	SVM	14	87
	LR	13	84.80
Kodati et al. (2015)	NB	13	83.70
Uyar et al. (2017)	GA based RFNN	13	96.63
Miao et al. (2016)	Adaptive boosting	29	80.14
Naidu et al. (2012)	K-mean based MAFIA with ID3	NA	85.0
Shamsollahi et al. (2019)	ANN	20	90.4
	C&RT	20	92.6
Santhanam et al. (2013)	PCA1+FFNN	8 ± 1	95.20
Latha & Jeeva (2019)	Majority vote with NB, BN, RF, and MP	9	85.48
Gupta et al. (2019)	MIFH	28	93.44
Our system	DT	5	98.9

Conclusion

The Healthcare sector is one of the largest and fast-growing sectors involving unprocessed complex patient's data. An intelligent model can be designed by incorporation of effective and efficient data. This data can help to make an intelligent model which can save human life. CVDs are life-threatening disease

across the globe and it is not limited to specific age factors. So, it is essential to diagnose it which can reduce the chances of premature death. In this article, we have presented a completely automated cardiovascular disease detection system that can detect the presence or absence of CVD.

Our main objective was not restricted to CVDs prediction, but also to figure out the most effective features that can describe the association between input features and target class. Additionally, a deeper understanding was required; why some features can carry the essential information to separate the presence or absence of heart disease and some others do not possess such capability. The majority of the prior predictive model of CVDs diagnosis uses 14, 20, 12, and 29 features in which a single re-sampling technique has been applied. Moreover, our feature selection approach and feature subset can improve the overall classification accuracy. The proposed system accuracy of 98.9% with an AUC value of 0.99 is better as compared to the prior diagnostic model.

Instead of using 14 features, 5 features can reduce the clinical testing expenses which help to reduce the patient financial burden. Moreover, such a computer-aided diagnostic system can work side by side with medical experts to provide a second opinion and actively assist CVDs detection of patients.

Limitations and Future Direction

Our proposed intelligent model is very promising but has some limitations due to the used classification algorithm. Overfitting is a most practical problem in the DT which is dependent on the size of the used dataset. For this reason, the structure of the tree is affected by small changes in data. We have dealt with 303 instances, but an increasing number of samples may create instability of a predefined tree model. On the other hand, the decision tree-based model requires less effort for understanding. Even nonlinear relationship cannot affect the decision tree structure and is considered as a helpful tool in case of multi-class problems. Apart from model limitations, we can apply the proposed approach along with other available heart disease datasets.

REFERENCES

Joshi, A., Rienks, M., Theofilatos, K., & Mayr, M. (2021). Systems biology in cardiovascular disease: a multiomics approach. *Nature Reviews Cardiology*, *18*(5), 313-330.

Tichenor, M., & Sridhar, D. (2019). Metric partnerships: global burden of disease estimates within the World Bank, the World Health Organisation and the Institute for Health Metrics and Evaluation. *Wellcome Open Research*, *4*.

Kasthuri A. (2018). Challenges to Healthcare in India - The Five A's. *Indian journal of community medicine : official publication of Indian Association of Preventive & Social Medicine*, *43*(3), 141–143. https://doi.org/10.4103/ijcm.IJCM_194_18

Peiris, D., Ghosh, A., Manne-Goehler, J., Jaacks, L. M., Theilmann, M., Marcus, M. E., ... & Geldsetzer, P. (2021). Cardiovascular disease risk profile and management practices in 45 low-income and middle-income countries: A cross-sectional study of nationally representative individual-level survey data. *PLoS medicine*, *18*(3), e1003485.

GRF Collaborators. (2018). A systematic analysis for the global burden of disease study 2017. *The Lancet*, *392*(10159), 1789-1858.

Jennum, P. J., Plazzi, G., Silvani, A., Surkin, L. A., & Dauvilliers, Y. (2021). Cardiovascular disorders in narcolepsy: Review of associations and determinants. *Sleep medicine reviews*, 101440.

Sabzmakan, L., Morowatisharifabad, M. A., Mohammadi, E., Mazloomy-Mahmoodabad, S. S., Rabiei, K., Naseri, M. H., ... & Mirzaei, M. (2014). Behavioral determinants of cardiovascular diseases risk factors: A qualitative directed content analysis. *ARYA atherosclerosis*, *10*(2), 71.

Abbara, S., Blanke, P., Maroules, C. D., Cheezum, M., Choi, A. D., Han, B. K., ... & Leipsic, J. (2016). SCCT guidelines for the performance and acquisition of coronary computed tomographic angiography: A report of the society of Cardiovascular Computed Tomography Guidelines Committee: Endorsed by the North American Society for Cardiovascular Imaging (NASCI). *Journal of cardiovascular computed tomography*, *10*(6), 435-449.

Haq, I. U., Haq, I., & Xu, B. (2021). Artificial intelligence in personalized cardiovascular medicine and cardiovascular imaging. *Cardiovascular Diagnosis and Therapy*, *11*(3), 911.

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future healthcare journal*, *6*(2), 94.

Louis, D. N., Gerber, G. K., Baron, J. M., Bry, L., Dighe, A. S., Getz, G., ... & Golden, J. A. (2014). Computational pathology: an emerging definition. *Archives of pathology & laboratory medicine*, *138*(9), 1133-1138.

Lagan, J., Schelbert, E. B., Naish, J. H., Vestbo, J., Fortune, C., Bradley, J., ... & Miller, C. A. (2021). Mechanisms underlying the association of chronic obstructive pulmonary disease with heart failure. *JACC: Cardiovascular Imaging*.

Dey, N., Ashour, A. S., Fong, S. J., & Bhatt, C. (Eds.). (2018). *Healthcare data analytics and management*. Academic Press.

Singh, Y. K., Sinha, N., & Singh, S. K. (2016, November). Heart disease prediction system using random forest. In *International Conference on Advances in Computing and Data Sciences* (pp. 613-623). Springer, Singapore.

Kausar, N., Palaniappan, S., Samir, B. B., Abdullah, A., & Dey, N. (2016). Systematic analysis of applied data mining based optimization algorithms in clinical attribute extraction and classification for diagnosis of cardiac patients. In *Applications of intelligent optimization in biology and medicine* (pp. 217-231). Springer, Cham.

Lysaght, T., Lim, H. Y., Xafis, V., & Ngiam, K. Y. (2019). AI-assisted decision-making in healthcare. *Asian Bioethics Review*, *11*(3), 299-314.

Dey, N., Ashour, A. S., Shi, F., Fong, S. J., & Sherratt, R. S. (2017). Developing residential wireless sensor networks for ECG healthcare monitoring. *IEEE Transactions on Consumer Electronics*, *63*(4), 442-449.

Suganya, R., Rajaram, S., Abdullah, A. S., & Rajendran, V. (2017). Prediction of heart diseases using hybrid feature selection and modified Laplacian pyramid non-linear diffusion with soft computing methods. International Journal of Biomedical Engineering and Technology, 25(1), 30. doi:10.1504/ijbet.2017.086550

Salhi, D. E., Tari, A., & Kechadi, M. T. (2020). Using Machine Learning for Heart Disease Prediction. In *CSA* (pp. 70-81).

Xiao-Yan Gao, Abdelmegeid Amin Ali, Hassan Shaban Hassan, Eman M. Anwar, "Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method", Complexity, vol. 2021, Article ID 6663455, 10 pages, 2021.

Gárate-Escamila, A. K., El Hassani, A. H., & Andrès, E. (2020). Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked, 19, 100330.

Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. Digital health, 6, 2055207620914777.

Almustafa, K.M. Prediction of heart disease and classifiers' sensitivity analysis. BMC Bioinformatics 21, 278 (2020)

Bashir, Saba, et al. "Improving heart disease prediction using feature selection approaches." 2019 16th international bhurban conference on applied sciences and technology (IBCAST). IEEE, 2019.

Youness Khourdifi and Mohamed Bahaji (2019). The Hybrid Machine Learning Model Based on Random Forest Optimized by PSO and ACO for Predicting Heart, ICCWCS 2019, April 24-25, Kenitra, Morocco.

Alotaibi, F. S. (2019). Implementation of Machine Learning Model to Predict Heart Failure Disease. International Journal of Advanced Computer Science and Applications (IJACSA), 10(6). http://dx.doi.org/10.14569/IJACSA.2019.0100637.

Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", Mobile Information Systems,vol. 2018, Article ID 3860146, 21 pages, 2018.

Kodati, S. "Analysis of Heart Disease using in Data Mining Tools Orange and Weka". Global Journal of Computer Science and Technology, vol 18-1, 2018.

Le, H. M., Tran, T. D., & Van Tran, L. A. N. G. (2018). Automatic heart disease prediction using feature selection and data mining technique. *Journal of Computer Science and Cybernetics*, *34*(1), 33-48.

Uyar, Kaan, and Ahmet Ilhan. Diagnosis of Heart Disease Using Genetic Algorithm Based Trained Recurrent Fuzzy Neural Networks. Procedia Computer Science, vol. 120, 2017, pp. 588–593.

Kumar, S. S., & Shaikh, T. (2017). Empirical Evaluation of the Performance of Feature Selection Approaches on Random Forest. 21017 International Conference on Computer and Applications (ICCA). doi:10.1109/comapp.2017.8079769

Pahwa, K., & Kumar, R. (2017). Prediction of heart disease using hybrid technique for selecting features. 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON). doi:10.1109/upcon.2017.8251100

Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2016). Prediction of heart disease using random forest and feature subset selection. In Innovations in bio-inspired computing and applications (pp. 187-196). Springer, Cham.

Dey, A., Singh, J., & Singh, N. (2016). Analysis of supervised machine learning algorithms for heart disease prediction with reduced number of attributes using principal component analysis. International Journal of Computer Applications, 140(2), 27-31.

Feshki, M. G., & Shijani, O. S. (2016). Improving the heart disease diagnosis by evolutionary algorithm of PSO and Feed Forward Neural Network. 2016 Artificial Intelligence and Robotics (IRANOPEN). doi:10.1109/rios.2016.7529489

Khan, S.S. 2016. Prediction of Angiographic Disease Status using Rule Based Data Mining Techniques. Biological Forum – An International Journal, 8(2), pp 103-107.

Khanna, D., et al. 2015. Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease. International Journal of Machine Learning and Computing, vol. 5, no. 5, pp. 414–419.

https://archive.ics.uci.edu/ml/datasets/heart+disease

Janecek, A., Gansterer, W., Demel, M., & Ecker, G. (2008, September). On the relationship between feature selection and classification accuracy. In *New challenges for feature selection in data mining and knowledge discovery* (pp. 90-105). PMLR.

Baranidharan, B., Pal, A., & Muruganandam, P. (2019). Cardiovascular disease prediction based on ensemble technique enhanced using extra tree classifier for feature selection. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(3), 3236-3242.

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational statistics & data analysis*, *52*(4), 2249-2260.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, *10*(1), 1-16.

Mythili, T., Mukherji, D., Padalia, N., & Naidu, A. (2013). A heart disease prediction model using SVM-Decision Trees-Logistic Regression (SDL). *International Journal of Computer Applications*, *68*(16).

Pandey, A. K., Pandey, P., Jaiswal, K. L., & Sen, A. K. (2013). A heart disease prediction model using decision tree. *IOSR Journal of Computer Engineering (IOSR-JCE)*, *12*(6), 83-86.

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, *11*(1), 1-13.

Palimkar, P., Shaw, R. N., & Ghosh, A. (2022). Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach. In *Advanced Computing and Intelligent Technologies* (pp. 219-244). Springer, Singapore.

Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., & Ghuli, P. (2020). Heart disease prediction using machine learning. *International Journal of Research and Technology*, *9*(04), 659-662.

Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In *2020 international conference on electrical and electronics engineering (ICE3)* (pp. 452-457). IEEE.

Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2020). HDPM: an effective heart disease prediction model for a clinical decision support system. *IEEE Access*, *8*, 133034-133050.

Ali, L., Wajahat, I., Golilarz, N. A., Keshtkar, F., & Bukhari, S. A. C. (2021). LDA–GA–SVM: improved hepatocellular carcinoma prediction through dimensionality reduction and genetically optimized support vector machine. *Neural Computing and Applications*, *33*(7), 2783-2792.

Miao, K. H., Miao, J. H., & Miao, G. J. (2016). Diagnosing coronary heart disease using ensemble machine learning. *International Journal of Advanced Computer Science and Applications*, *7*(10), 1-12.

Naidu, P. M., & Rajendra, C. (2012). Detection of Health Care using Data Mining Concepts Through Web. *International Journal of Advanced Research in Computer Engineering and Technology*, *1*(4), 45-50.

Shamosollahi, M, Badiee, A., & Ghazanfari M. (2019). Using Combined Descriptive and Predictive Methods of Data Mining for Coronary Artery Disease Prediction: a Case Study Approach. Journal of Artificial Intelligence & Data Mining (JAIDM). 7(1), 47-58.

Santhanam, T., & Ephzibah, E. P. (2013). Heart disease classification using PCA and feed forward neural networks. In *Mining Intelligence and Knowledge Exploration* (pp. 90-99). Springer, Cham.

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203.

Gupta, A., Kumar, R., Arora, H. S., & Raman, B. (2019). MIFH: A machine intelligence framework for heart disease diagnosis. *IEEE Access*, *8*, 14659-14674.

Kausar, N., Abdullah, A., Samir, B. B., Palaniappan, S., AlGhamdi, B. S., & Dey, N. (2016). Ensemble clustering algorithm with supervised classification of clinical data for early diagnosis of coronary artery disease. *Journal of Medical Imaging and Health Informatics*, *6*(1), 78-87.