# Early-Stage Cardiovascular Disease Prediction Using a Sigmoidtropy-Based Decision Tree

Anurag Bhatt* and Ashutosh Kumar Bhatt

*Uttarakhand Open University, Haldwani, India*

**Abstract:** Heart disease (HD) is a significant health issue in the world, and its early and proper prediction is essential to minimize mortality and the development of the disease. Cardiovascular disease (CVD) is one of the diseases that need effective and stable predictive models to assist clinical decision-making. This paper gives a Sigmoidtropy-Based Decision Tree (SDT) model of cardiovascular disease prediction, which improves the traditional decision tree by adding a sigmoid-based formulation of entropy. The heart disease data are first grouped by the K-means clustering method in order to enhance the data representation. The suggested SDT model is tested on the Cleveland heart disease dataset of the UCI repository and compared to the traditional classifiers, such as Naive bayes, random forest, and the traditional Decision Tree models. Experimental findings indicate that the SDT has an accuracy of 99.67 which is better than the performance of Random Forest (76.89%), Decision Tree (76.56%), and Naive Bayes (81.84%) with a lower execution time. Despite the promising performance shown by the results, it needs further validation with more datasets and strong evaluation plans to determine the generalizability.

**Keywords**: Cardiovascular Disease, Sigmoidtropy-Based Decision Tree (SDT), K-Means Clustering, Naive Bayes, Random Forest, Machine Learning, Risk Prediction, Data Mining.

## 1. INTRODUCTION

Cardiovascular disease (CVD) is a highly common and dangerous health disorder in the entire world and a cause of the greatest mortality rate. The development of CVD is in most instances very fast and therefore there is little time to intervene clinically unless it is identified at a tender age. As a result, the health care fraternity has a big problem of properly identifying patients within a reasonable time. Misdiagnosis or late diagnosis does not only influence patient outcomes, but also has an effect on the credibility and operational efficiency of health institutions. Moreover, the treatment of CVD is quite expensive, and in the developing world, including India, a significant percentage of patients cannot afford the long-term treatment [1], [2]. The rising global mortality related to heart related diseases over the past years has shown the necessity of having reliable, accurate and cost-effective predictive models that can help in making early diagnosis and timely treatment.

As healthcare data continues to expand at an alarming rate, sophisticated computing methods are now necessary to analyze large and complex medical data. Deep learning and machine learning methods have been extensively used to automate the process of finding knowledge and decision making in healthcare applications. Specifically, Naive Bayes, Support Vector Machines (SVM), Decision Trees (DT), K-Nearest Neighbour (KNN) and Random Forest (RF) are some of the most commonly used supervised learning models to predict heart diseases and clinical decision support system [3, 4]. The models are useful because they help healthcare professionals to increase the accuracy of the diagnosis and minimize the reliance on manual analysis.

Early diagnosis of cardiovascular disease is important in determining high-risk individuals, particularly the ones over the age of 30, so that preventive methods can be taken against the condition like lifestyle change, medical counselling, and, medication in time before the disease advances to a higher level. Nevertheless, the current predictive methods are usually unable to work with incomplete, noisy, or poorly formatted clinical data, which have a negative impact on the model performance. Poor management of missing data and past patient data can result in inaccurate forecasts and restrict the performance of early disease diagnosis.

Despite promising outcomes with conventional machine learning models, most of the currently existing methods use conventional measures of entropy or information gain in decision tree models. These approaches can have a low sensitivity to complex feature distributions, marginal entropy differences, or skewed data, and can make suboptimal choices of splits and have low predictive robustness. Additionally, various works concentrate mainly on the methods of feature selection or ensemble learning, whereas relatively little effort has been devoted to the improvement of the split evaluation mechanism of decision tree models per se. This observation creates a gap in the research on identifying alternative entropy formulations that can enhance the performance of decision trees in predicting cardiovascular disease.

In order to overcome these shortcomings, this paper suggests a Sigmoidtropy-Based Decision Tree model, where the entropy measure is transformed by a

*Address correspondence to this author at the Uttarakhand Open University, Haldwani, India; E-mail: anurag15bhatt@gmail.com

sigmoid to increase the split discrimination and stability when constructing a tree. The suggested approach will enhance the accuracy of classification and retain computational efficiency and interpretability, which is why it is applicable in clinical decision-support settings.

The key contributions of the proposed algorithm of Sigmoidtropy-Based Decision Tree are as follows: To handle both structured and unstructured data available in datasets of cardiovascular diseases so as to enhance the performance of prediction.

- To automatically extract informative features out of structured clinical data, in consultation with health care professionals, to increase the predictive accuracy.
- To create a powerful cardiovascular disease risk prediction model using the appropriate clinical attributes.
- To show, by experimental assessment, that the proposed Sigmoidtropy-Based Decision Tree is empirically better than the current state-of-the-art procedures.

The rest of this paper is structured in the following way: Section 2 will provide a literature review of related literature. Section 3 explains the research methodology that will be used including the proposed algorithm. Section 4 is the discussion of experimental results and comparison. Lastly, Section 5 wraps up the paper and gives future research directions.

## 2. LITERATURE REVIEW

Machine learning (ML) methods have been actively implemented to predict cardiovascular disease (CVD) and heart disease (HD) because they can process complex clinical data and aid in making medical decisions. Initial predictive models were mainly centred on supervised learning methods to categorize patients with or without cardiovascular disease, and experimental validation was usually done using Python-based analytical settings to evaluate the accuracy and reliability of the algorithms [5].

Analytical frameworks that are based at the system level and network level have also been discussed to determine cardiovascular risk, especially in patients with comorbidities. This was a disease-network-based ML model that was proposed to predict cardiovascular risk in patients with type-2 diabetes where disease networks were built based on cohort-based data and network-derived features were utilized to train various ML models [6]. The accuracy of the prediction reported was between 79% and 88% which indicates the potential of network analytics and machine learning. Simultaneously, a survey-based study was conducted to compare the use of supervised and unsupervised learning methods, such as ANN, DT, Fuzzy Logic, KNN, Naïve Bayes, SVM, and Logistic Regression, and give a systematic review of their suitability and drawbacks in heart disease prediction tasks [7].

The use of feature selection has been identified as one of the determinants of predictive performance in cardiovascular disease models. Some of them explored how to determine meaningful clinical characteristics and use them with the appropriate classifiers. The classification frameworks that were based on voting had a precision of about 87.4 percent when they were used on optimized feature subsets [8]. Equally, feature identification methods that employed machine-learning achieved a maximum accuracy of 88.7% on heart disease prediction models [9]. Also, more general analytical literature investigated the use of ML in echocardiography, electrocardiography, and more advanced non-invasive imaging modalities, as well as the issues associated with interpretability, data heterogeneity, and clinical adoption [10].

Ensemble and hybrid learning techniques have been extensively used in order to increase predictive accuracy. Hybrid models were developed based on decision trees with artificial neural networks and proved to be more accurate, sensitive, and specific, especially when benchmark datasets of the UCI repository were used to validate them [11]. Hybrid frameworks based on feature-selection further supported the performance improvement of classes of classifiers, like DT, Logistic Regression, SVM, Random Forest (RF), and Naive Bayes, using tools such as RapidMiner [12]. Other decision-tree improvement methods, including splitting based on Gini-index and discretization methods, were also demonstrated to be more effective in prediction accuracy and sensitivity than the traditional tree-based approaches [13].

Comparative studies have always shown that the performance of the Random Forest models is very strong especially in missing data and larger data sets. According to one study, RF had high sensitivity, specificity, precision, and area-under-the-curve (AUC) of 94.7 percent on the UCI heart disease data [14]. Hybrid learning approaches also showed that the combination of classifiers can be used to enhance predictive performance compared to the performance of single models [15]. Naive Bayes classifiers were also found to work better when used with feature-selection methods like recursive feature elimination and gain-ratio methods [16]. Further assessments with the Cleveland data set have indicated 83.49 percent accuracy when all the 13 clinical attributes are taken into account [17, 18].

Specific studies on decision-tree-based methods showed that there were significant performance differences across algorithmic settings. In some prediction tasks, Standalone Decision Tree models were reported to have a 77.55% accuracy [19]. When the techniques of boosting were used on Decision Trees, performance was improved and the accuracy was greater compared to basic implementations of DTs [20]. Applications of the J48 algorithm achieved an accuracy of 67.7% indicating gradual increase over the previous methods [21]. More complex settings, including alternating decision tree with the principal component analysis, had a greater accuracy level of up to 92.2% [22]. Classifiers based on decision trees that included forward feature selection also have been reported to have better weighted accuracy [24].

Random Forest approaches based on ensembles proved to be robust in a variety of datasets. According to one of the studies, RF had an accuracy of 91.6% on the Cleveland dataset and 97% on the People's Hospital dataset [25]. A different study was found to have an F-measure of 0.86 with RF-based classification to predict cardiovascular disease [26]. The prediction of coronary heart disease with the help of the Random Forest app also demonstrated the accuracy of 97.7, which supports the efficiency of the ensemble learning methods [27].

In addition to structured clinical data, disease prediction with the use of ML has been applied to unstructured and multimodal healthcare data. The social-media data were analyzed using fuzzy association-rule-based methods to examine trends associated with healthcare and forecast possible risks to health [28]. Medical image analysis, including brain tumor detection, was successfully implemented with deep learning methods which spurs the growth of further studies on the application of advanced ML methods to cardiovascular imaging and early detection of sudden cardiac events [29].

Recent studies have paid more attention to sophisticated machine learning systems, explainable artificial intelligence (XAI), and combined optimization techniques. Angiographic-based weighted SVM models showed better diagnostic performance with optimized parameter selection [30]. Smart healthcare systems using RF, DT, and KNN as an ensemble further improved the reliability of prediction of early heart diseases detection [31]. ML models based on feature-selection and hybrid models were always reported to be strong in classification across various datasets [32]. Holoistic ensemble designs with integrated machine learning and deep learning models had better sensitivity and specificity [33]. The superior hybrid optimization-based models minimized overfitting and enhanced predictive accuracy, as compared to the traditional classifiers [34].

Recent cardiovascular prediction studies have also made interpretability and transparency important considerations. Population-based datasets were used to construct interpretable ML models based on SHAP analysis to determine essential predictors of coronary heart disease [35]. The explicable ensemble-learning models also showed better accuracy and model transparency on various datasets [36]. Benchmarking experiments have established that boosting and bagging ensembles are superior predictors compared to the traditional classifiers [37]. ML frameworks based on interpretable and IoMT showed encouraging outcomes in real-time cardiovascular monitoring and prediction [38, 39]. Further studies confirmed the application of predictive analytics and data-mining methods in the early detection of cardiovascular risks and real-time observations in a variety of clinical practices [40, 41].

In order to offer a systematic analysis of the current methods, Table **1** is a synthesis of representative studies, datasets, methodologies, and reported metrics of performance in cardiovascular disease prediction.

**Table 1: Comparative Summary of Existing Machine Learning Approaches for Cardiovascular Disease Prediction**

| Ref. | Dataset | Method(s) Used | Key Contribution | Reported Performance |
|---|---|---|---|---|
| [6] | Australian cohort (T2D) | Network analytics + ML | Disease-network-based risk modeling | Accuracy: 79–88% |
| [8] | Clinical HD dataset | KNN, DT, NB, SVM, Vote | Feature selection + voting | Accuracy: 87.4% |
| [9] | Clinical HD dataset | HRFLM + ML | Feature identification | Accuracy: 88.7% |
| [11] | UCI HD | DT + ANN (Hybrid) | Improved sensitivity & specificity | Improved over single models |
| [13] | UCI HD | Gini-based DT | Alternative split criterion | Improved precision |
| [14] | UCI HD | NB, SVM, DT, LR, RF | Robustness to missing data | AUC: 94.7% |
| [22] | UCI HD | Alternating DT + PCA | Feature reduction | Accuracy: 92.2% |
| [25] | UCI HD | Random Forest | Ensemble robustness | Accuracy: 91.6% |
| [27] | Clinical dataset | Random Forest | Coronary HD prediction | Accuracy: 97.7% |
| [36] | Multi-dataset | Ensemble + XAI | Explainability + accuracy | Improved transparency |

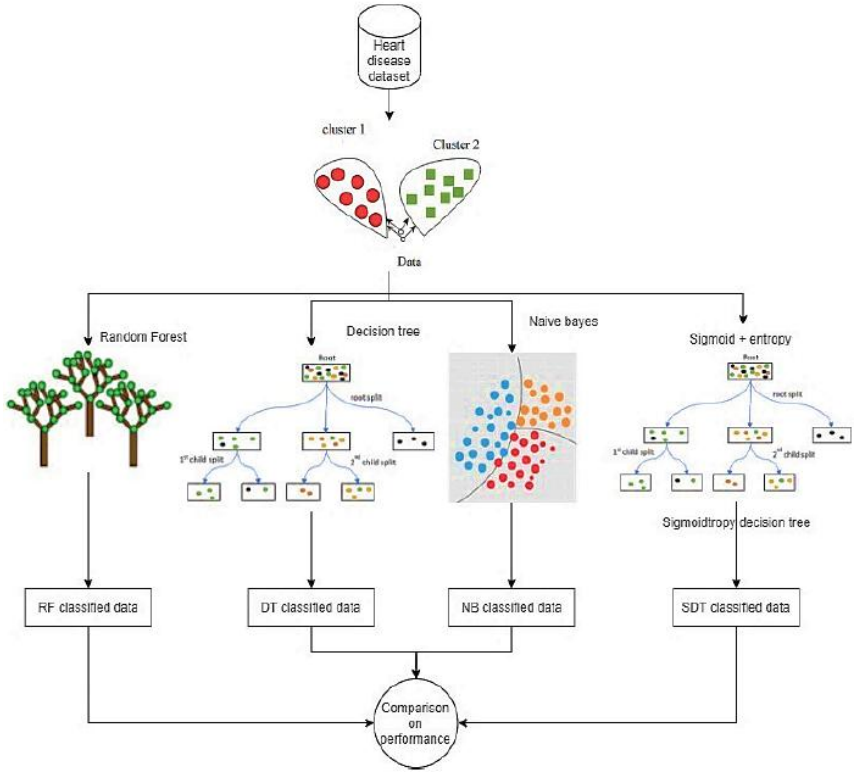



**Figure 1:** Experiment workflow with CVD Dataset.

## 3. RESEARCH METHODOLOGY

### 3.1. Overview of the Proposed Methodology

The proposed methodology is based on a systematic approach to the prediction of cardiovascular disease at an early stage as shown in Figure **1**. It starts with the selection of pertinent clinical features of the UCI heart disease data. The first step is to cluster these features by a clustering method to group similar patterns of data. The clustered data are then fed into several classification models such as Naive Bayes, Decision Tree, Random Forest and the proposed Sigmoidtropy-Based Decision Tree (SDT). This workflow aims at assessing the possibility of optimizing decision-tree-based prediction by modifying the entropy-based split criterion.

### 3.2. Dataset and Feature Description

This study uses the dataset provided by the UCI Machine Learning Repository and it includes the most popular clinical features related to the prediction of heart disease. There are 13 features that are taken into

**Table 2: Features in Dataset**

| S.No | Heart Disease Dataset Parameter | | |
|---|---|---|---|
| 1 | Age | Numeric | No. of years |
| 2 | Gender | Numeric | Patient gender |
| 3 | Cp | Numeric | Pain class |
| 4 | Trestbps | Numeric | Blood pressure in resting state (mm Hg) |
| 5 | Fbs | Numeric | Blood sugar in fasting state (mg/dl) |
| 6 | Chol | Numeric | Cholesterol of serum (mg/dl) |
| 7 | Thalach | Numeric | Heart rate |
| 8 | Restecg | Numeric | ECG pattern |
| 9 | Slope | Numeric | Peak exercise ST segment slope |
| 10 | Exang | Numeric | Angina due to exercise |
| 11 | Ca | Numeric | No. of fluoroscopy colored vessel |
| 12 | Old peak | Numeric | Rest relative - exercise induced ST depression |
| 13 | Thal | Numeric | Status of defect |

account such as demographic data, physiological measurements and the results of diagnostic tests. Table **2** summarizes these features. In preprocessing, records that contained missing values were processed before analysis and numerical attributes were normalized so that they were scaled equally before clustering and classification.

### 3.3. K-Means Clustering of Feature Data

First, the characteristics added to the heart disease data are clustered with the help of the sequential K-means algorithm. Clustering is meant to cluster together feature values that share similar attributes before classification hence grouping the data structure before learning algorithms are applied. According to the predetermined cluster size K, the data is separated into K clusters. Euclidean distance between the data points and the cluster centroid is computed and cluster means are updated repeatedly till stable clusters are achieved. In this analysis, K = 2 is used to indicate the binary nature of heart disease classification (presence or absence of disease).

**Algorithm 1: Clustering**

1. Initialize the selected feature dataset
2. Divide the dataset based on cluster size $K$
3. Estimate the mean of each subset
4. Estimate the Euclidean distance for each division
5. Form clusters based on mean and Euclidean distance
6. Recompute the mean for each cluster
7. Compare and update each cluster's average value
8. Produce the final clustered dataset

### 3.4. Classification Models

Following the clustering, the data is then provided to three baseline classifiers, including Naïve Bayes, Decision Tree and Random Forest, to predict heart disease. Such models are applied to set up comparative performance standards with the proposed approach.

Naive Bayes is a probabilistic classifier that relies on the Bayes theorem and assumes that predictor variables are independent. Nevertheless, this assumption notwithstanding, it is popular because of its simplicity and computing efficiency.

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)} \quad (1)$$

Where

$P(c \mid x)$is the posterior probability,

$P(x \mid c)$is the likelihood,

$P(c)$is the class prior probability, and

$P(x)$is the predictor prior probability.

$$P(c \mid X) = P(x_1 \mid c)P(x_2 \mid c) \ldots P(x_n \mid c)P(c) \quad (2)$$

Decision Tree is a supervised learning algorithm that recursively divides the dataset according to the values of the attributes. Entropy and information gain are used to determine the splitting attributes.

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \quad (3)$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (4)$$

Random Forest is a supervised learning algorithm that is an ensemble-based algorithm which builds a number of decision trees and identifies the final class label by majority voting. The method enhances generalization and decreases overfitting.

### 3.5. Proposed Sigmoidtropy-Based Decision Tree

The suggested Sigmoidtropy-Based Decision Tree is a modification of the traditional decision tree which implements a sigmoid transformation of entropy values applied in split evaluation. This change constrains the entropy values and changes the sensitivity of split-selection process.

The Sigmoidtropy is given by the following function:

$$\sigma(S) = \frac{1}{1 + e^{-S}}, \sigma(S_v) = \frac{1}{1 + e^{-S_v}} \quad (5)$$

Using Sigmoidtropy, the modified information gain is computed as:

$$Gain(S, A) = \sigma(S) - \sum_{v \in Values(A)} \frac{\mid S_v \mid}{\mid S \mid} \sigma(S_v) \quad (6)$$

**Algorithm 2: Sigmoidtropy-Based Decision Tree**

1. Generate the root node $N$
2. Determine the class distribution for each attribute
3. If all instances belong to the same class, assign that class to node $N$
4. If attribute values are null, assign the majority class to node $N$

5. Select the attribute with the highest Sigmoidtropy-based gain ratio
6. Split the dataset based on the selected attribute
7. Recursively construct subtrees
8. Terminate when stopping conditions are met

## 4. RESULTS AND DISCUSSION

### 4.1. Experimental Setup

The proposed Sigmoidtropy-Based Decision Tree (SDT) model is applied to overview its performance. The environment of the doctor node and user node is built on a 64-bit Intel Core processor of the frequency of 2.45 GHz. Java is also used in transaction level prototyping. The suggested SDT framework is applied and the current classifiers, i.e. Decision Tree (DT), Naive Bayes (NB), and Random Forest (RF) are used to allow comparing the performance of the classifiers in the same experimental setting.

### 4.2. Dataset Description and Evaluation Metrics

#### *4.2.1. Dataset Selection*

The cardiovascular disease data is acquired on the UCI Machine Learning Repository. The database used in this work is the Cleveland database which is one of the four databases available on heart diseases. The data is made up of 303 cases and 14 attributes.

In the proposed model, thirteen attributes are used to analyze performance, with the age attribute not included, which is in line with the experimental design as described in the methodology.

#### *4.2.2. Performance Evaluation Metrics*

The proposed SDT and the existing classifiers are built into a confusion matrix. Accuracy, precision, recall, F-measure, and ROC area are used to evaluate the performance of the classifiers. Also, the error based measures such as Kappa statistic, Mean Absolute Error (MAE), Root Mean square error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared error (RRSE) are applied to evaluate prediction error and classification agreement.

### 4.3. Performance analysis

#### *4.3.1. Feature Clustering Results*

The heart disease data provided by UCI has over thirteen attributes, and they are grouped together through K-means clustering algorithm. Figure **2a-m** demonstrates the visual representation of the data of clustered features. The features of the clustering process are grouping the values similar to each other, and the number of clusters is established to K=2. The two clusters reflect in-control and out-of-control feature values, which are further applied to classification.

#### 4.3.2. Error-Based Performance Metrics

This section of the research paper explains the findings of the confusion matrix and error measures of different classifiers, which are Random Forest, Naive Bayes, Decision Tree, and Sigmoidtropy-Based Decision Tree. Table **3** shows these classifier performance in terms of Kappa statistic, MAE, RMSE, RAE, RRSE and execution time.

For the Random Forest classifier, the Kappa statistic, MAE, RMSE, RAE, and RRSE are recorded as 0.5235, 0.2617, 0.389, 54.11%, and 75.58%, respectively, with a response time of 618 ms. The Decision Tree classifier achieves Kappa, MAE, RMSE, RAE, and RRSE values of 0.5141, 0.2597, 0.4549, 53.69%, and 88.40%, respectively, with an execution time of 227 ms. For the Naïve Bayes classifier, the corresponding values are 0.6281, 0.2023, 0.3794, 41.83%, and 73.72%, with a time consumption of 299 ms.

The proposed Sigmoidtropy-Based Decision Tree achieves Kappa, MAE, RMSE, RAE, and RRSE values of 0.9934, 0.0066, 0.0576, 1.35%, and 11.20%, respectively, with a reduced execution time of 85 ms. The comparative behavior of these error metrics across classifiers is illustrated in Figure **3a-e**.

#### *4.3.3 Computational Time Analysis*

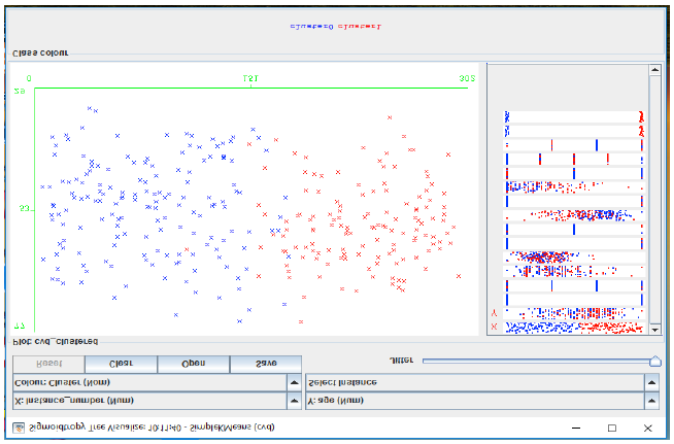
Figure **4** shows comparison of the execution time of the classifiers. The average time of execution of Random Forest, Decision Tree, Naive Bayes, and the proposed Sigmoidtropy-Based Decision Tree is noted as 618 ms, 227 ms, 299 ms, and 85 ms, respectively. These findings show that the proposed SDT takes less computational time as opposed to the other classifiers in the same experimental set-up.

#### *4.3.4. Classification Performance Based on Confusion Matrix*

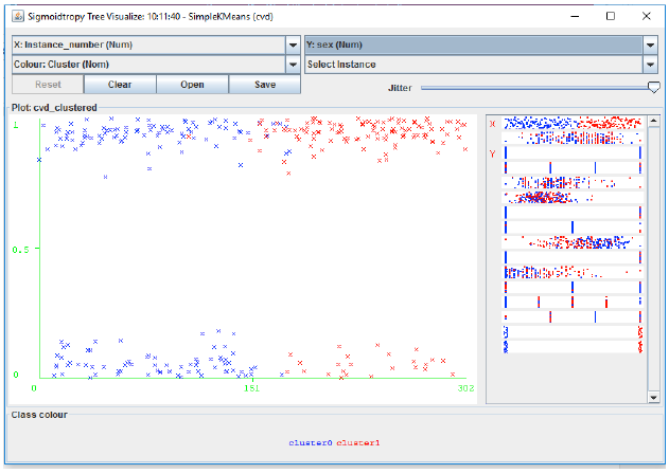
The classification performance based on confusion-matrix-derived metrics is summarized in Table **4**, which reports precision, recall, F-measure, ROC area, and accuracy for all classifiers. The accuracy values of Random Forest, Decision Tree, and Naïve Bayes are 76.89%, 76.56%, and 81.84%, respectively. The proposed Sigmoidtropy-Based Decision Tree achieves an accuracy of 99.67%, along with high precision, recall, and F-measure values.

The relative performance of precision, recall, and F-measure of each of the classifiers is also demonstrated in Figure **5**, whereas the ROC curve of the suggested SDT model is presented in Figure **6**.

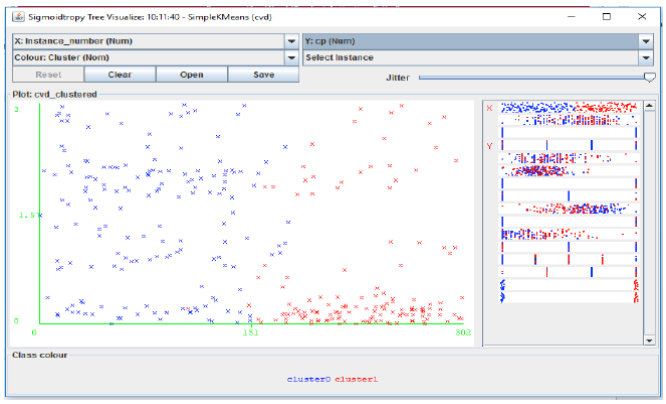
Despite the high accuracy value of the proposed SDT, the findings are derived with one benchmark dataset and thus should be viewed with caution.



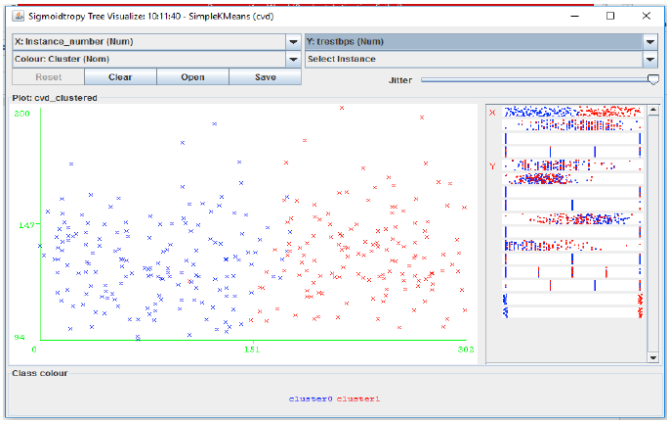
(**a**)


Sigmoidtropy Tree Visualizer: 10:11:40 - SimpleKMeans (cvd)
X: instance_number (Num)
Y: sex (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: cvd_clustered
Class colour
cluster0 cluster1

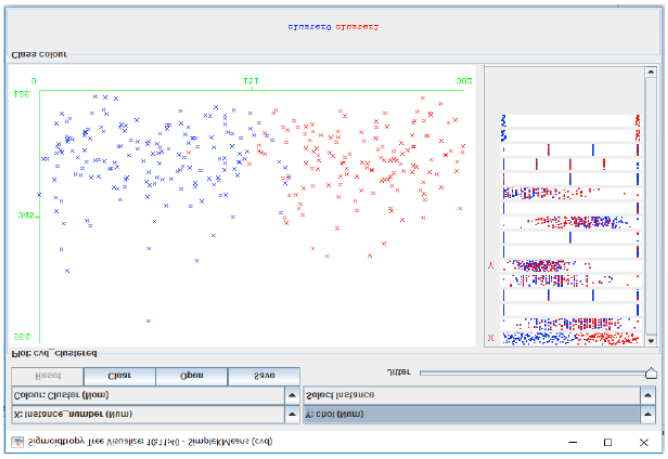
(**b**)


Sigmoidtropy Tree Visualizer: 10:11:40 - SimpleKMeans (cvd)
X: instance_number (Num)
Y: cp (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: cvd_clustered
Class colour
cluster0 cluster1

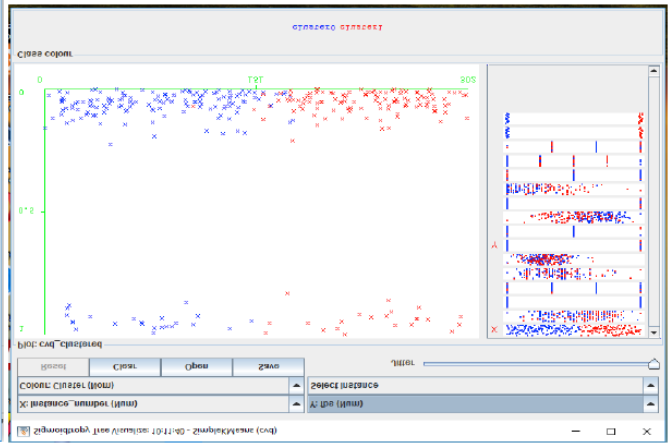
(**c**)


Sigmoidtropy Tree Visualizer: 10:11:40 - SimpleKMeans (cvd)
X: instance_number (Num)
Y: trestbps (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: cvd_clustered
Class colour
cluster0 cluster1

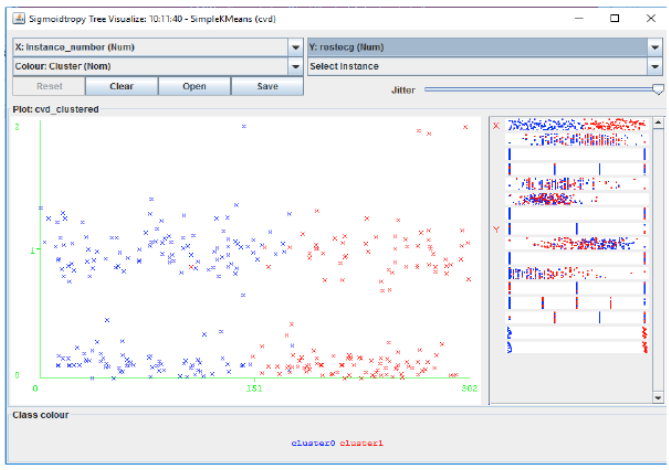
(**d**)



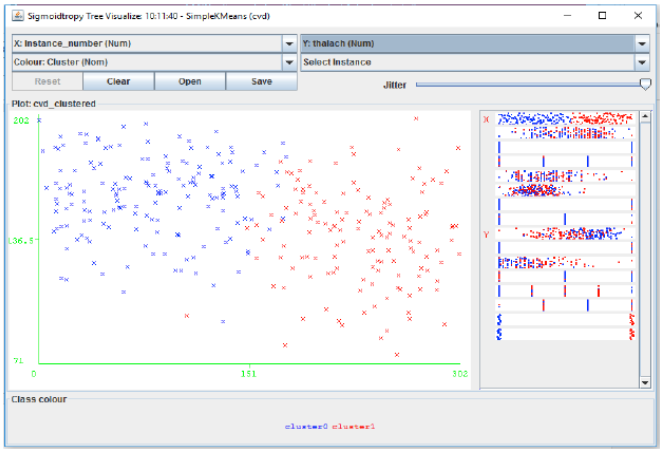
(**e**)



(**f**)


Sigmoidtropy Tree Visualizer: 10:11:40 - SimpleKMeans (cvd)
X: instance_number (Num)
Y: restecg (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: cvd_clustered
Class colour
cluster0 cluster1

(**g**)


Sigmoidtropy Tree Visualizer: 10:11:40 - SimpleKMeans (cvd)
X: instance_number (Num)
Y: thalach (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: cvd_clustered
Class colour
cluster0 cluster1

(**h**)

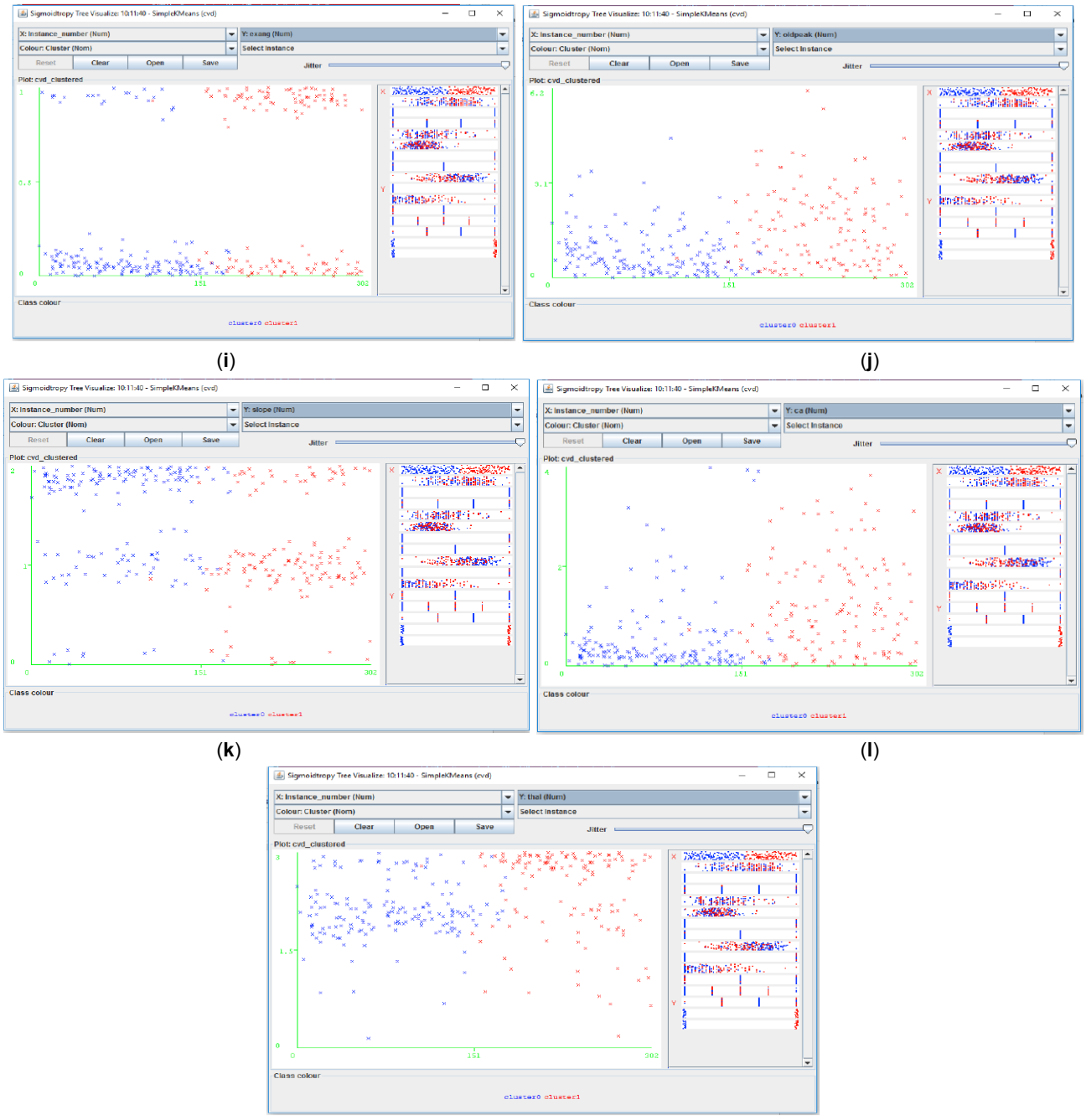



**Figure 2:** (**a**). K-means clustering result for Feature 1 of the heart disease dataset. (**b**). K-means clustering result for Feature 2 of the heart disease dataset. (**c**). K-means clustering result for Feature 3 of the heart disease dataset. (**d**). K-means clustering result for Feature 4 of the heart disease dataset. (**e**). K-means clustering result for Feature 5 of the heart disease dataset. (**f**). K-means clustering result for Feature 6 of the heart disease dataset. (**g**). K-means clustering result for Feature 7 of the heart disease dataset. (**h**). K-means clustering result for Feature 8 of the heart disease dataset. (**i**). K-means clustering result for Feature 9 of the heart disease dataset. (**j**). K-means clustering result for Feature 10 of the heart disease dataset. (**k**). K-means clustering result for Feature 11 of the heart disease dataset. (**l**). K-means clustering result for Feature 12 of the heart disease dataset. (**m**). K-means clustering result for Feature 13 of the heart disease dataset.

**Table 3: Performance on Algorithms on Error Metrics**

| Description | Random forest | Decision tree | Naïve Bayes | Sigmoidtropy tree |
|---|---|---|---|---|
| K-Stat (Kappa statistic) | 0.5235 | 0.5141 | 0.6281 | 0.9934 |
| MAE (Mean absolute error) | 0.2617 | 0.2597 | 0.2023 | 0.0066 |
| RMSE (Root mean squared error) | 0.389 | 0.4549 | 0.3794 | 0.0576 |
| RAE (Relative absolute error) | 0.5411 | 0.5369 | 0.4183 | 0.0135 |
| RRSE (Root relative squared error) | 0.7558 | 0.8840 | 0.7372 | 0.1120 |
| Time Utilized (in ms) | 618 | 227 | 299 | 85 |

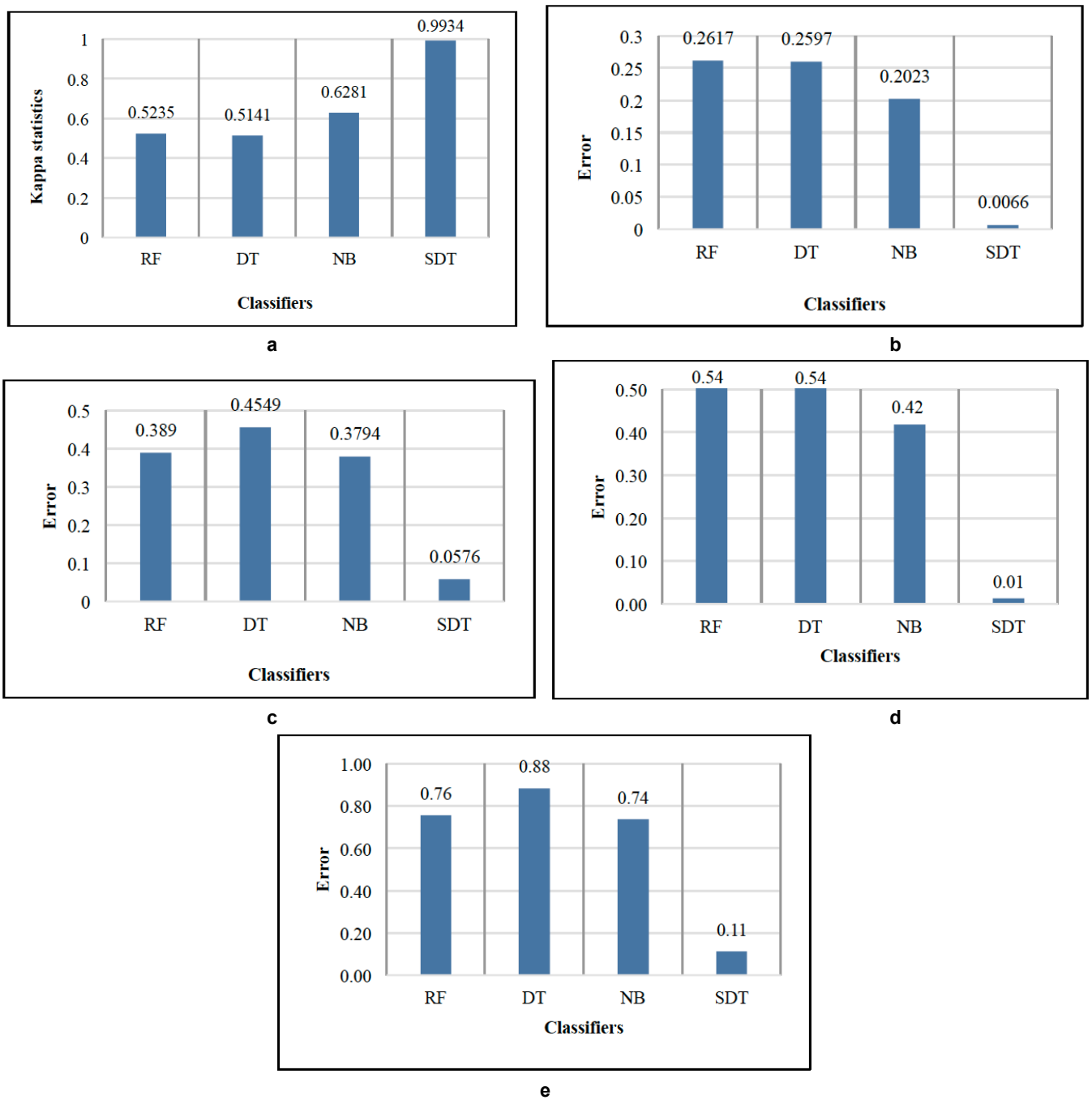



**Figure 3: a.** Comparison of Kappa statistic values for different classifiers. **b.** Comparison of mean absolute error values for different classifiers. **c.** Comparison of root mean square error values for different classifiers. **d.** Comparison of relative absolute error values for different classifiers. **e.** Comparison of root relative squared error values for different classifiers.

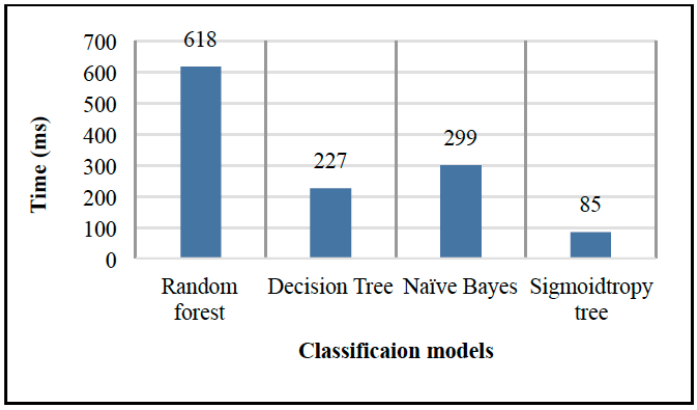



**Figure 4:** Comparison of execution time for Random Forest, Decision Tree, Naïve Bayes, and Sigmoidtropy-Based Decision Tree classifiers.

**Table 4: Performance Based on Confusion Matrix**

| Algorithm | Precision | Recall | F-measure | ROC -area | Accuracy |
|---|---|---|---|---|---|
| Random Forest | 0.782 | 0.769 | 0.762 | 0.864 | 76.89% |
| DT | 0.788 | 0.766 | 0.756 | 0.74 | 76.56% |
| NB | 0.826 | 0.818 | 0.816 | 0.896 | 81.84% |
| SDT | 0.997 | 0.997 | 0.997 | 0.993 | 99.67% |

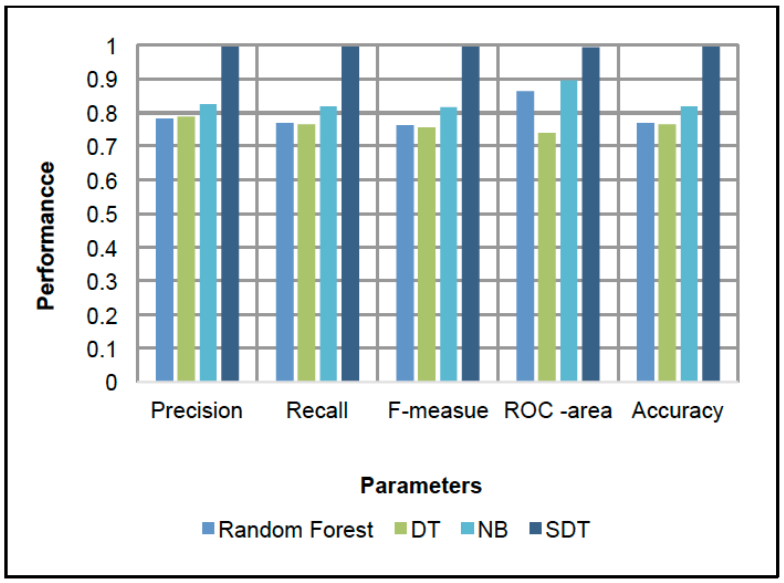



**Figure 5:** Comparison of precision, recall, and F-measure values for different classifiers.

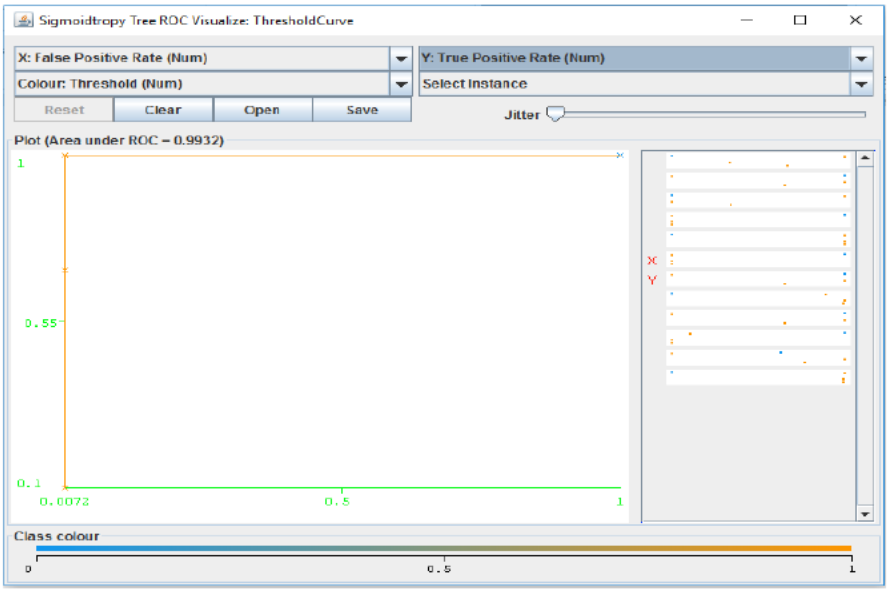



**Figure 6:** Receiver operating characteristic (ROC) curve of the Sigmoidtropy-Based Decision Tree classifier.

Further validation methods like cross-validation or testing on external data would be needed to further test the performance of generalization.

## 5. DISCUSSION

The experimental findings suggest that addition of a sigmoid-based entropy transformation in the decision tree learning procedure affects the classification accuracy in various measures of evaluation. The proposed Sigmoidtropy-Based Decision Tree has been shown to perform better in terms of error minimization, classification accuracy and execution time as compared to the traditional classifiers.

Application wise, smaller prediction error and less computational cost are good attributes of

decision-support system in the healthcare setting. Nevertheless, the results of this research are restricted to the experiment conditions and data set. More research based on more data and statistical validation methods is needed to determine the strength and applicability of the suggested method.

## 6. CONCLUSION

The proposed Sigmoidtropy-Based Decision Tree is intended to determine significant features using machine-learning methods to enhance the accuracy of prediction of cardiovascular disease (CVD). To determine the efficiency of the proposed approach, the forecasting framework is tested with different combinations of features and a number of popular classification algorithms, such as Naive Bayes, Decision Tree, and Random Forest. As shown in the experiments, the accuracy of the Random Forest, Decision Tree, Naïve Bayes and Sigmoidtropy-Based Decision Tree classifiers is 76.89, 76.56, 81.84 and 99.67, respectively. The computed performance of the Random Forest, Decision Tree, Naive Bayes, and the proposed Sigmoidtropy-Based Decision Tree classifiers is 618 ms, 227 ms, 299 ms, and 85 ms, respectively, in terms of computational efficiency. The validation findings show that the proposed model is more accurate in classification and has a low computation time than the available classifiers in the same experimental conditions. These results indicate that a sigmoid-based entropy adjustment can be used to improve decision-tree learning to predict CVD. Nevertheless, the findings are derived based on a single benchmark dataset and thus need to be viewed with caution. Additional validation with various data sets, cross-validation methods and statistical significance analysis is needed to determine robustness and generalization. Future research will be aimed at the expansion of the suggested methodology to structured and unstructured CVD data and the investigation of more sophisticated data-mining and analytics tools to aid in the reliable clinical decision-making process and enhance cardiovascular healthcare outcomes.

## REFERENCES


[1] Taneja A, Heart disease prediction system using data mining techniques. Oriental Journal of Computer science and Technology 2013; 6(4): 457-66.

[2] Gawali M, Shirwalkar N, Kalshetti A. Heart disease prediction system using data mining techniques. International Journal of Pure and Applied Mathematics 2018; 120(6): 499-506.

[3] Iliyas MM, Shaikh MI, Student MC. Prediction of Heart Disease Using Decision Tree. Allana Inst of Management Sciences, Pune 2019; 9: 1-5.

[4] Yang L, Wu H, Jin X, Zheng P, Hu S, Xu X, Yu W, Yan J. Study of cardiovascular disease prediction model based on random forest in eastern China. Scientific reports 2020; 10(1): 1-8. https://doi.org/10.1038/s41598-020-62133-5

[5] Chauhan YJ. Cardiovascular Disease Prediction using Classification Algorithms of Machine Learning. International Journal of Science and Research (IJSR) 2020; 9(5): 194-200.

[6] Hossain ME, Uddin S, Khan A. Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes. Expert Systems with Applications 2021; 164: 113918. https://doi.org/10.1016/j.eswa.2020.113918

[7] Singh R. A Review on Heart Disease Prediction using Unsupervised and Supervised Learning. Neural Networks 99: 100.

[8] Amin MS, Chiam YK, Varathan KD. Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics 2019; 36: 82-93. https://doi.org/10.1016/j.tele.2018.11.007

[9] Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access 2019; 7: 81542-54. https://doi.org/10.1109/ACCESS.2019.2923707

[10] Al'Aref SJ, Anchouche K, Singh G, Slomka PJ, Kolli KK, Kumar A, Pandey M, Maliakal G, Van Rosendael AR, Beecy AN, Berman DS. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. European heart journal 2019; 40(24): 1975-86. https://doi.org/10.1093/eurheartj/ehy404

[11] Maji S, Arora S. Decision tree algorithms for prediction of heart disease. In Information and communication technology for competitive strategies. Springer, Singapore 2019; pp. 447-454. https://doi.org/10.1007/978-981-13-0586-3_45

[12] Bashir S, Khan ZS, Khan FH, Anjum A, Bashir K. Improving heart disease prediction using feature selection approaches. In2019 16th international bhurban conference on applied sciences and technology (IBCAST). IEEE 2019; pp. 619-623. https://doi.org/10.1109/IBCAST.2019.8667106

[13] Mathan K, Kumar PM, Panchatcharam P, Manogaran G, Varadharajan R. A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease. Design automation for embedded systems 2018; 22(3): 225-42. https://doi.org/10.1007/s10617-018-9205-4

[14] Li R, Shen S, Zhang X, Li R, Wang S, Zhou B, Wang Z. Cardiovascular Disease Risk Prediction Based on Random Forest. In The International Conference on Healthcare Science and Engineering. Springer, Singapore 2018; pp. 31-43. https://doi.org/10.1007/978-981-13-6837-0_3

[15] Esfahani HA, Ghazanfari M. Cardiovascular disease detection using a new ensemble classifier. In 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI). IEEE 2017; pp. 1011-1014. https://doi.org/10.1109/KBEI.2017.8324946

[16] Pahwa K, Kumar R. Prediction of heart disease using hybrid technique for selecting features. In2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON). IEEE 2017; pp. 500-504. https://doi.org/10.1109/UPCON.2017.8251100

[17] Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In 2017 IEEE symposium on computers and communications (ISCC). IEEE 2017; pp 204-207. https://doi.org/10.1109/ISCC.2017.8024530

[18] Cleveland Heart Disease Dataset, [online] Available: http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[19] Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease.. In 2017 IEEE Symposium on Computers and Communications (ISCC). IEEE 2017; pp 205-206. https://doi.org/10.1109/ISCC.2017.8024530

[20] Bouali H, Akaichi J. Comparative study of different classification techniques: heart disease use case. In2014 13th International Conference on Machine Learning and Applications. IEEE 2014; pp. 482-486. https://doi.org/10.1109/ICMLA.2014.84

[21] Ekız S, Erdoğmuş P. Comparative study of heart disease classification. In2017 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT). IEEE 2017; pp. 1-4. https://doi.org/10.1109/EBBT.2017.7956761

[22] Chauhan R, Bajaj P, Choudhary K, Gigras Y. Framework to predict health diseases using attribute selection mechanism. In2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom). IEEE 2015; pp. 1880-1884.

[23] Jabbar MA, Deekshatulu BL, Chndra P. Alternating decision trees for early diagnosis of heart disease. InInternational Conference on Circuits, Communication, Control and Computing. IEEE 2014; pp. 322-328. https://doi.org/10.1109/CIMCA.2014.7057816

[24] Farooq K, Karasek J, Atassi H, Hussain A, Yang P, MacRae C, Mahmud M, Luo B, Slack W. A novel cardiovascular decision support framework for effective clinical risk assessment. In2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE). IEEE 2014; pp. 117-124. https://doi.org/10.1109/CICARE.2014.7007843

[25] Xu S, Zhang Z, Wang D, Hu J, Duan X, Zhu T. Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework. In2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA). IEEE 2017; pp. 228-232. https://doi.org/10.1109/ICBDA.2017.8078813

[26] Rahman QA, Tereshchenko LG, Kongkatong M, Abraham T, Abraham MR, Shatkay H. Utilizing ECG-based heartbeat classification for hypertrophic cardiomyopathy identification. IEEE transactions on nanobioscience 2015; 14(5): 505-12. https://doi.org/10.1109/TNB.2015.2426213

[27] Shahin A, Moudani W, Chakik F, Khalil M. Data mining in healthcare information systems: case studies in Northern Lebanon. InThe Third International Conference on e-Technologies and Networks for Development (ICeND2014). IEEE 2014; pp. 151-155. https://doi.org/10.1109/ICeND.2014.6991370

[28] Mittal M, Kaur I, Pandey SC, Verma A, Goyal LM. Opinion Mining for the Tweets in Healthcare Sector using Fuzzy Association Rule. EAI Endorsed Transactions on Pervasive Health and Technology 2019; 4(16). https://doi.org/10.4108/eai.13-7-2018.159861

[29] Mittal M, Arora M, Pandey T, Goyal LM. Image segmentation using deep learning techniques in medical images. In Advancement of Machine Intelligence in Interactive Medical Image Analysis. Springer, Singapore 2020; pp. 41-63. https://doi.org/10.1007/978-981-15-1100-4_3

[30] Banjoko AW, Abdulazeez KO. Efficient data-mining algorithm for predicting heart disease based on an angiographic test. Malays J Med Sci 2021; 28(5): 118-129. https://doi.org/10.21315/mjms2021.28.5.12

[31] Absar N, Das EK, Shoma SN, Uddin M, *et al*. The efficacy of machine-learning-supported smart system for heart disease prediction. Healthcare (Basel) 2022; 10(6): 1137. https://doi.org/10.3390/healthcare10061137

[32] Biswas N, Ali MM, Rahaman MA, Islam M, Mia MR, Azam S, Ahmed K, Bui FM, Al-Zahrani FA, Moni MA. Machine learning-based model to predict heart disease in early stage employing different feature selection techniques. BioMed Res Int 2023; 2023: 6864343. https://doi.org/10.1155/2023/6864343

[33] Sadr H, Salari A, Ashoobi MT, Nazari M, *et al*. Cardiovascular Disease Diagnosis: a holistic approach using the integration of machine learning and deep learning models. Eur J Med Res 2024; 29: 455. https://doi.org/10.1186/s40001-024-02044-7

[34] Rehman MU, Naseem S, Butt AR, Mahmood T, Khan AR, Khan I, Khan J, Jung YH, *et al*. Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment. Sci Rep 2025; 15: 13361. https://doi.org/10.1038/s41598-025-96437-1

[35] Vu T, Kokubo Y, Inoue M, Yamamoto M, Mohsen A, Martin-Morales A, *et al*. Machine learning model for predicting coronary heart disease risk: development and validation using insights from a Japanese population-based study. JMIR Cardio 2025; 9: e68066. https://doi.org/10.2196/68066

[36] Ganie SM, Pramanik PKD, Zhao Z. Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets. Sci Rep 2025; 15(1): 13912. https://doi.org/10.1038/s41598-025-97547-6

[37] Teja MD, Rayalu GM. Optimizing heart disease diagnosis with advanced machine learning models: a comparison of predictive performance. BMC Cardiovasc Disord 2025; 25: 212. https://doi.org/10.1186/s12872-025-04627-6

[38] Kailasanathan N, *et al*. Heart disease prediction with a feature-sensitized interpretable framework for Internet of Medical Things sensors. Front Digit Health 2025. https://doi.org/10.3389/fdgth.2025.1612915

[39] Bhatt A, Dubey S, Bhatt A. Sudden cardiac arrest prediction using predictive analytics. Int J Intell Eng Sys 2017; 10(3): 184-191. https://doi.org/10.22266/ijies2017.0630.20

[40] Bhatt A, Dubey SK, Bhatt AK. Early prediction of cardiovascular disease among young adults through coronary artery calcium score technique. In: International Conference on Advances in Computing and Data Sciences; . Cham: Springer International Publishing 2021; pp. 303-312. https://doi.org/10.1007/978-3-030-88244-0_29

[41] Bhatt A, Bhatt AK. Leveraging data mining for predictive insights into cardiovascular health risks. Journal of Applied Bioanalysis 2025; 11(S2): 137-146. https://doi.org/10.53555/jab.v11si2.516

Received on 07-11-2025 Accepted on 12-12-2025 Published on 30-12-2025

https://doi.org/10.6000/1929-6029.2025.14.77