Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/120121

Performance Analysis Of Various Machine Learning Techniques To Predict Cardiovascular Disease: An Emprical Study

M. Chandralekha[*] *and N. Shenbagavadivu*

Department of Computer Applications, University College of Engineering, Bharathidasan Institute of Technology (BIT) Campus, Anna University, Tiruchirappalli - 620 024, India

Received: 28 Sep. 2017, Revised: 6 Dec. 2017, Accepted: 10 Dec. 2017
Published online: 1 Jan. 2018

Abstract: In the modern state-of-art of technology, Machine Learning emerges out as a boom to extract information from mammoth dataset and transform into acquainted information. In particular, Clustering (Unsupervised learning) and Classification (Supervised learning) are the two predominant Machine Learning approaches emphasized here. However, data and constraints are known primarily in Classification, they are unknown in Clustering. In recent times, Clustering and Classification started playing significant role in the area of innumerable applications like Cognitive Services, Image Recognition and Manipulation, Business and Legal, Text and Language, Medical, Weather Forecast, Genetics, Bio-informatics and so on. A few recently established machine learning methodologies are depicted here, with a provision to convey vital concepts to classification and clustering experts. The aim of this paper is to focus various Machine Learning techniques through which one can predict the heart disease of a patient by analyzing various medical diagnostic parameters and patterns. A comparative study is made with respect to both unsupervised learning (Partitioning-based, Hierarchical-based, Density-based and Model-based clustering) and supervised learning (SVM, Random Forest (RF), Decision tree (DT) and K-nn) empirically with the inclusion of large number of datasets. The results are explicit that Decision Tree has more classification accuracy of 73% thereby correlating K-means, K-modes, K-medoids, CLARANS, PAM, FCM, CLARA, DBSCAN, Ward's, ROCK, FCM, SVM, EM, OPTICS, Random Forest and K-nn. In this perspective, R X64 3.1.3 is used as a tool to determine the accuracy of aforementioned algorithms.

Keywords: Machine Learning, Classification,Clustering, K-means, K-medoids, K-modes, PAM, CLARANS, CLARA, FCM, Ward's, ROCK, DBSCAN, OPTICS, EM, SVM, Decision tree, Random Forest, K-nn, R X64 3.1.3

1 Introduction

Nowadays almost all hospitals started maintaining the large amount of data in e-form to generate their own medical details. The quantum of data is getting accrued by and large day by day as the hospitals handle different forms of records which include both structured and unstructured data like images, texts, values etc. These data are extensively useful to tap knowledgeable information such as pattern generation enabling Knowledge Discovery in Databases (KDD).

Machine Learning techniques are widely used to detect heart diseases by employing the University of California Irvine (UCI) heart disease dataset also known as the Cleveland dataset[1]-[5]. Few researchers focused different Machine Learning techniques using different data sets to achieve highest accuracy [6]-[13]. Previous related researches are enlisted beneath:

By conducting 10x10 fold-cross-validation , the experimental results for predicting heart disease using SVM, Multi-Layer Perceptron (MLP), C4.5, Bayesian classifiers methodologies were attained by Jovic et al. [14] with the set of sensitivity / specificity percentile as 77.2/87.4, 96.6/97.8, 99.2/98.4 and 98.4/99.2 respectively.

Using Classification and Regression Tree(CART) with feature selection technique Mellilo et al. [15] reported Sensitivity as 89.74% and Specificity as 100.00% by 10 fold-cross-validations.

Yu et al. [16] has adopted Feature selection by four different (UCIMFS, MIFS, CMIFS, mRMRb) SVM

[*] Corresponding author e-mail: rdbmchandralekha@gmail.com

approaches to evaluate the following measures with their respective percentage (Sensitivity: 96.55, 93.10, 93.10, 93.10; Specificity: 98.14,98.14, 100.00, 98.14; Accuracy: 97.59, 96.38,97.59, 96.38) thereby using Leave-one-out cross-validations. Liu et al. [17] analyzed the results and revealed that by applying Feature selection, Feature normalization and Feature combination of SVM & k-NN techniques with the evaluation measures as cent percentage accuracy, precision and sensitivity in SVM whereas 91.49 % accuracy, 94.12 % precision, 84.21 % sensitivity in KNN using Cross validation.

It was previously observed by Narin et al. [18] through incorporating the Filter based backward elimination feature selection of SVM, k-NN, LDA, MLP, RBF classifier with the following set of results as 82.75 % of sensitivity, 96.29 % of Specificity, 91.56 % of Accuracy in SVM; 65.51 % of sensitivity, 96.29 % of Specificity, 85.54 % of Accuracy in k-NN; 75.86 % of sensitivity, 90.74 % of Specificity, 85.54 % of Accuracy in Polynomial LDA; 82.75 % of sensitivity, 92.59 % of Specificity, 89.15 % of Accuracy in MLP; 58.62 % of sensitivity, 96.29 % of Specificity, 93.13 % of Accuracy in RBF by Leave-One-Out cross-validation.

Zheng et al. [19] has stated that through Least Square Support Vector Machine (LS-SVM) method in which Accuracy of 95.39 %, Sensitivity of 96.59 % and Specificity of 93.75 % were obtained by double-fold cross-validation.

Masetic et al. [20] has provided the Random Forest method in an attempt to achieve cent percent in ROC area, F-measure and accuracy thereby using 10-fold cross-validation

Bohacik et al. [21] obtained the accuracy of 77.66%, Sensitivity of 37.31% and Specificity of 91.53% by using Alternating decision tree technique with 10-fold cross-validation.

This paper aims towards concluding the most efficient technique among K-means, K-medoids, K-modes, PAM, CLARANS, CLARA, FCM, Ward's, ROCK, DBSCAN, OPTICS, EM, SVM, Decision tree, Random Forest and K-nn employed for the prediction of heart disease on the basis of accuracy or prediction rate using R X64 3.1.3 machine learning software.

Subsequently, this paper consists of Section 2 in which categories of machine learning algorithms are reviewed in detail. Section 3 is dealing with empirical study of machine learning algorithms. Section 4 focusses performance analysis of different machine learning algorithms. Section 5 highlights the proposed research. Finally, Section 6 concludes with the findings and performance efficacy.

2 Categorization of Machine Learning Techniques

This section illustrates the broad categorization of machine learning algorithms comprising both Unsupervised learning (Clustering) based on Partition, Hierarchy, Density, Model and supervised learning (Classification) to compare, analyse & establish their merits and demerits. Figure 1 provides categorization of machine learning approaches.

Unsupervised learning is used to learn appropriate structure without labelled classes or any other information beyond the raw data. One of the widely used unsupervised learning methods is cluster analysis in which an extensive research is made to generate unknown patterns or core structure in data. Accordingly, Clusters are depicted with the help of similarity or distance measures like Cosine similarity, Euclidean distance, Kernel functions, Language modelling etc. Fahad et al. [22] elaborated numerous clustering algorithms. Some areas of applications are sequence and pattern mining in data mining, Bootstrapping Classification, Disambiguation, Machine Translation, Dependency Parsing , Morphology, Image segmentation in medical imaging , sequence analysis and genetic clustering in bioinformatics, object recognition in computer vision, Sentence & Word Segmentation etc. Several clustering algorithms are classified based on partitioning, hierarchical, density, grid and model. Discussions pertaining to the above mentioned algorithms are described as follows.

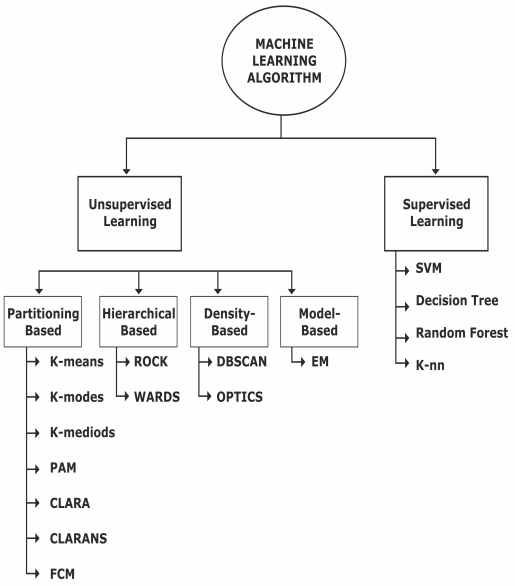

Fig. 1: Categorization of Machine Learning approaches.

Partitioning-based clustering algorithm: Finding clusters in partitioning-based clustering algorithm is an instant phenomenon. The purpose of partitioning algorithm is to split data objects into various partitions in which each partition denotes a cluster using a series of iterations. Every cluster must have different groups in which every group should have minimum of one object and every individual object should fit precisely in a group. In addition, few partitioning algorithms viz K-means, K-medoids, K-modes, Partitioning around Medoids (PAM),

Clustering Large Applications based upon RANdomized Search (CLARANS), Clustering LARge Applications (CLARA), Fuzzy c-means (FCM) are analysed.
Merits: Fewer time complexity and more computation efficiency.
Demerits: Inappropriate for concave data, partial response to noise (outliers).

Hierarchical-based clustering algorithm: Unlike partitioning-based clustering, hierarchy is followed in Hierarchical-based clustering in which proximity is used thereby considering orientation of nodes to form the tree like nested clusters so called Dendogram. There are two types of hierarchical-based clustering namely agglomerative (bottom-up) and divisive (top-down). It is initiated with single object in an agglomerative clustering for every cluster and iteratively combines more than two relevant clusters. Similarly, in a divisive clustering, initiation is carried out primarily by means of using the dataset as single cluster and iteratively separates the best suitable cluster. The above process takes place till it terminates. The bottleneck of this process is that it cannot be continued as and when it is getting combined or separated. Further, hierarchical algorithms ROCK(RObust Clustering using linKs) and Ward's are analysed vividly.
Merit: Highly scalable.
Demerit: More time complexity.

Density-based clustering algorithm: In Density-based clustering, data objects are scattered mainly with respect to the density pertaining to the points that are connected in density nearby, thus developing in any angle. The main features of this algorithm are to determine arbitrary shapes and noise handling. Also, Density-based clustering algorithms DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and OPTICS (Ordering points to identify the clustering structure) are examined here.
Merits: Higher efficiency, appropriate data with arbitrary shape.
Demerits: Low quality, requirement of huge memory space.

Model-based clustering algorithm: Augmentation of model-based algorithm is to suit the data and previously defined mathematical model. Assuming the data generation through combination of Probability Distributions thereby enabling to determine the quantum of clusters automatically using statistical standards by accounting noise (outliers) leading to strong clustering method. One of the model-based algorithms expectation-maximization (EM) is taken for investigation.
Merits: Different and advanced models provision to describe the data effectively.
Demerit: More time complexity.

In **Supervised learning** by using known dataset termed as labelled training dataset which includes both input data and response values and thereby generating a model to make predictions of the response values for a target data. As a result, validation of the above generated model is done through testing dataset. Various applications of the supervised machine learning algorithms include biometric attendance or ATM, spam filters, weather prediction, predicting winning % between two teams, Face detection, Text and speech categorization, Signature recognition and Medicine. Supervised learning includes two categories of algorithm:

1. Classification: for categorical response values, where the data can be separated into specific "classes".
2. Regression: for continuous-response values.

Common classification and regression algorithms analysed here include: Support vector machines (SVM), Decision tree, Nearest neighbors (kNN) and Random forest

3 Empirical study of Machine Learning Algorithms

In this section, a detailed investigation is carried out for finding the characteristics of various machine learning algorithms and compared the results achieved by analysing empirically to predict heart disease using South African Heart-Disease Dataset.

Subsequently, datasets are being described in sub-section 3 A, attributes are being tabulated in sub-section 3 B and metric analysis is being made in sub-section 3 C.

A. Dataset Description The large datasets used for this study is drawn from Rossouw et al. [23] considering the survey sample of males collected from risk prone region i.e. Western Cape of South Africa to detect the heart disease of the individuals indicating negative (0) or positive (1) . The positive result holders were asked to undergo remedial steps including blood pressure reduction to bring down the level of risk during post treatment.

B. Attribute Description The list of attributes in South African Heart-disease dataset is depicted in Table 1.

C. Metrics Analysis in Different Machine Learning Algorithms A powerful statistical, free and open source software tool R X64 3.1.3 is used to analyze data pertaining to various Machine Learning Algorithms viz K-means, K-medoids, K-modes, PAM, CLARANS, CLARA, FCM, Ward's, ROCK, DBSCAN, OPTICS, EM, SVM, Decision tree (DT), Random Forest (RF) and K-nn. Clusters in any number can be considered for analysis of the algorithms as shown in the Figure 2.

Table 1: Sample Dataset

ATTRIBUTE	DOMAIN
Sbp	systolic blood pressure[101,218]
Tobacco	cumulative tobacco (kg)[0.0,31.2]
Ldl	low density lipoprotein cholesterol[0.98, 15.33]
Adiposity	[6.74, 42.49]
Famhist	family history of heart disease {Present, Absent}
Typea	type-A behavior[13, 78]
Obesity	[14.7, 46.58]
Alcohol	current alcohol consumption[0.0, 147.19]
Age	age at onset[15, 64]
Chd	{0, 1}

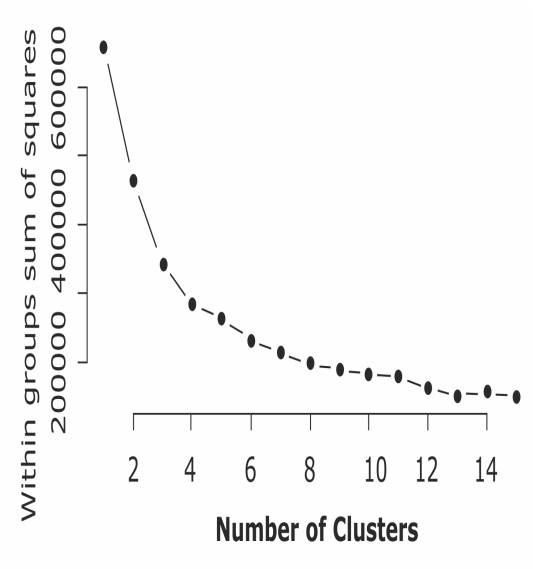

Fig. 2: Cluster traceability curve

The main objective of this study is to analyze the performance of different machine learning algorithms using R software package. The standard way of measuring the performance of machine learning algorithms is by calculating the precision and recall. Here, 60% of data collected are used for training while 40% is used for testing. Using the training and testing sets prepared, experiments were conducted on South African Heart-disease dataset. The selected machine learning algorithms extract patterns and build a model. The goal of building a model is to train and predict the different patterns available in the datasets. Thus whenever a new instance is provided to the model, the model predicts the instance to a particular class, the instances belong. The models are built for sixteen machine learning algorithms chosen for this study using the training set. The performance and the accuracy of prediction of the models are evaluated using different metrics. The results are shown in the following tables. Table 2 shows different metrics applied over the dataset for training whereas Table 3 shows different metrics applied over the dataset for testing. The performance and accuracy of these models differ largely on the dataset, size, features, instances and no of classes. Also, these models accuracy and performance suffer greatly with respect to missing values and would require a data pre-processing before applying the data to the model.

Another method of reckoning the performance of the algorithms is to use confusion matrix containing information pertaining to predicted (columns) and actual class (rows) which can be visualized easily. The representation of confusion matrix is shown in Figure 3.

		Predicted Class	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Fig. 3: Confusion Matrix

Confusion matrices obtained for some of the machine learning techniques such as K-means, K-medoids, K-modes, PAM, CLARANS, CLARA, FCM, Ward's, ROCK, DBSCAN, OPTICS and EM are discussed in Table 4.

In addition, a pictorial representation can also be performed using R X64 3.1.3 with machine learning algorithms such as Ward's, DBSCAN, OPTICS, Decision tree (DT) and Random Forest (RF) by taking into consideration of both training and testing data for South-African heart disease dataset produced are plotted and visualized as follows.

Ward's Dendogram produced after testing the data is shown in Figure 4.

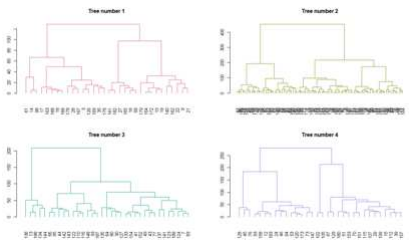

Fig. 4: Ward's Dendogram -Testing data

Ward's Dendogram produced after training the data is shown in Figure 5.

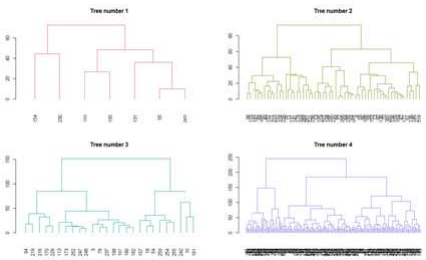

Fig. 5: Ward's Dendogram -Training data

DBSCAN cluster plot for training data is shown in Figure 6.

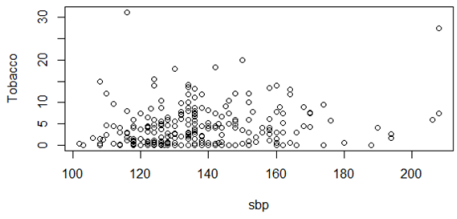

Fig. 6: DBSCAN -Training data

DBSCAN cluster for testing data is shown in Figure 7.

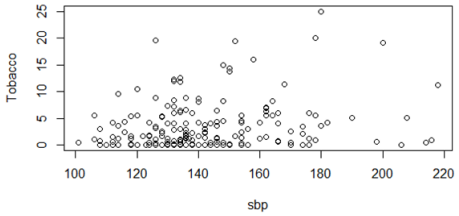

Fig. 7: DBSCAN -Testing data

OPTICS cluster plot for training data is shown in Figure 8.

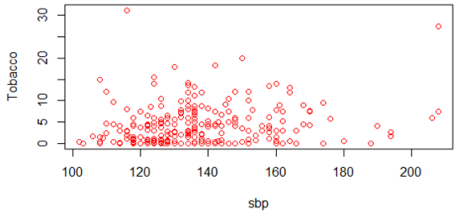

Fig. 8: OPTICS-Training data

OPTICS cluster for testing data is shown in Figure 9.

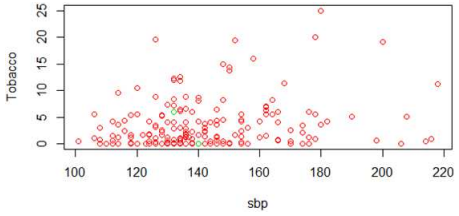

Fig. 9: OPTICS- Testing data

Complexity Parameter vs Relative error for Decision Tree produced after training the data is shown in Figure 10

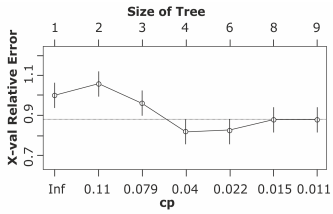

Fig. 10: Complexity Parameter vs Relative error

Decision Tree for training data is shown in Figure 11.

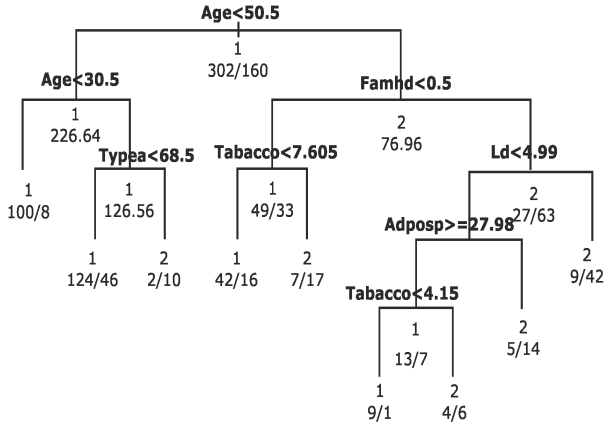

Fig. 11: Decision Tree-Training data

Pruned Decision Tree for training data is shown in Figure 12.

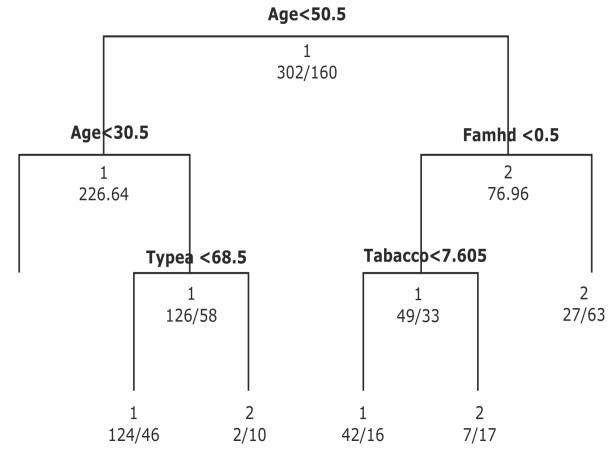

Fig. 12: Pruned Decision Tree

The precision vs recall plot is produced after training and testing in Random Forest is shown in Figure 13.

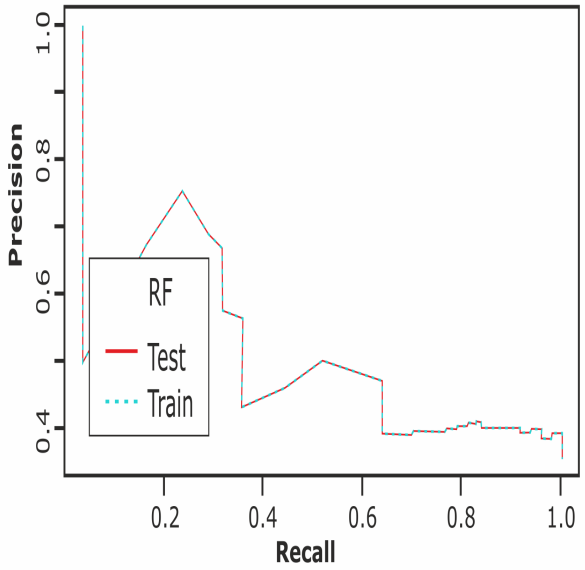

Fig. 13: Precision/Recall graph of Random Forest (Test/Train)

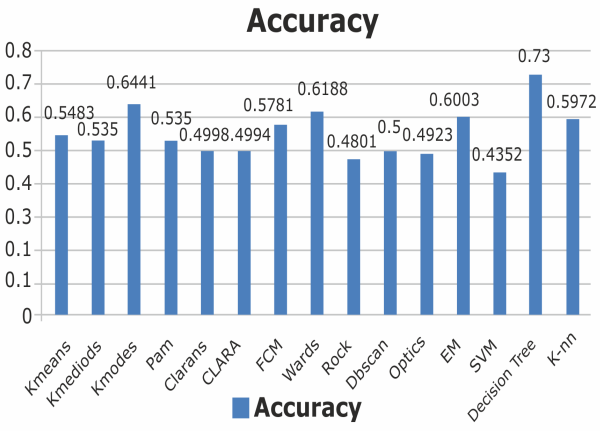

Fig. 14: Accuracy of Machine Learning algorithms

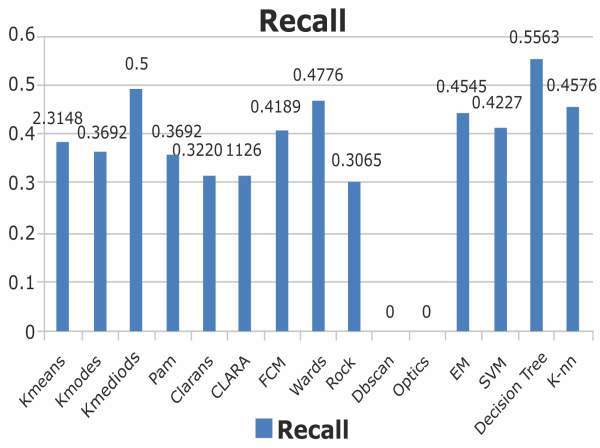

Fig. 15: Recall of Machine Learning algorithms

4 Performance Analysis

This section presents a comparison of various machine learning algorithms and the performance is interpreted using sensitivity, specificity and accuracy for South-African heart disease dataset. Table 5 gives the algorithm and its corresponding output values attained during experimental evaluation.

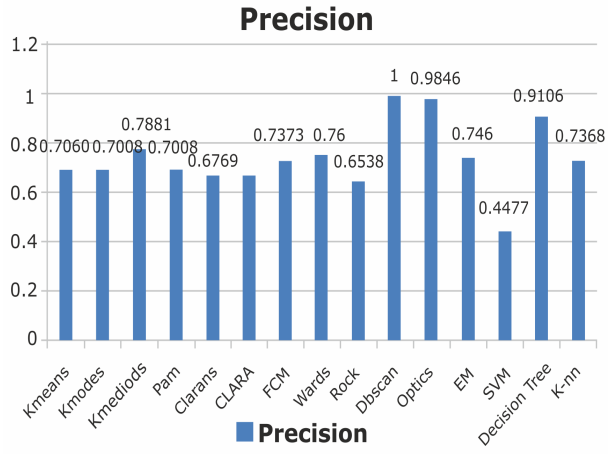

Fig. 16: Precision of Machine Learning algorithms

The final output of different classification results are plotted to compare the performance of the algorithm's on heart disease classification. Figure 14 reveals that DT has the highest classification accuracy of 0.73, while k-modes has achieved 64%, similarly Ward's hierarchal clustering and EM has 60% of accuracy. K-means, k-medoids, PAM, DBSCAN, and KNN has 50% and above accuracy. Sensitivity refers to correctly classified cases, i.e. classifying the cases that have heart disease. In our dataset we have 302 cases of positive heart disease out of total 462 cases.

According to Figure 15 Decision tree and k-modes have 55% and 50% of correctly classified instances. FCM, Ward's, EM, SVM and KNN have recall value of less than 50%.

Precision is the measure of classifying correctly out of all the available positive cases. From Figure 16 most of the algorithms have good precision values, i.e. the ability to classify the positive cases positively. DBSCAN has a higher precision value of 100% and OPTICS has 98% and Decision tree has 91%. K-means, K-medoids, K-modes, PAM, FCM, Ward's, EM and KNN have 70% and above.

5 Discussion and Future Work

The proposed method involves feature selection using information theory. Since wrapper methods rely on the accuracy of the classifier, they can be modified to fit any model based learning methods. On the other hand subsets created from wrapper method cannot affirm

Table 2: Statistical measures for training set

MEASURES	K means	K medoids	K modes	PAM	CLARANS	CLARA	FCM	WARDS	ROCK	DBSCAN	OPTICS	EM	SVM	DTree
Kappa	0.24	0.19	-0.10	0.19	-0.19	-0.25	0.18	0.03	-0.02	0.00	0.00	-0.35	-0.12	0.50
Mcnemar's P-Value	0.25	0.92	0.73	0.92	0.00	0.04	0.10	0.00	0.01	0.00	0.00	0.64	0.00	0.00
Sensitivity	0.72	0.71	0.60	0.71	0.53	0.52	0.71	0.65	0.52	1.00	1.00	0.52	0.45	0.91
Specificity	0.53	0.48	0.29	0.48	0.27	0.23	0.46	0.57	0.46	0.00	0.00	0.12	0.42	0.56
PosPred Value	0.77	0.72	0.62	0.72	0.38	0.44	0.64	0.98	0.63	0.64	0.64	0.54	0.58	0.79
NegPred Value	0.46	0.47	0.28	0.47	0.40	0.29	0.55	0.04	0.35	NaN	NaN	0.11	0.30	0.77
Prevalence	0.68	0.65	0.66	0.65	0.46	0.54	0.57	0.97	0.64	0.64	0.64	0.67	0.64	0.65
Detection Rate	0.49	0.46	0.40	0.46	0.25	0.28	0.41	0.63	0.33	0.64	0.64	0.35	0.29	0.60
Detection Prevalence	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.52	1.00	1.00	0.64	0.49	0.75
Balanced Accuracy	0.62	0.60	0.45	0.60	0.40	0.38	0.59	0.61	0.49	0.50	0.50	0.32	0.44	0.73

Table 3: Statistical measures for testing set

MEASURES	K means	K medoids	K modes	PAM	CLARANS	CLARA	FCM	WARDS	ROCK	DBSCAN	OPTICS	EM	kNN
Kappa	0.095	0.071	0.297	0.071	-0.001	-0.001	0.161	0.242	-0.039	0.000	-0.021	0.204	0.192
Mcnemar's P-Value	0.817	0.822	0.162	0.822	1.000	1.000	0.201	0.620	0.915	9.408	0.000	0.716	0.807
Sensitivity	0.707	0.701	0.788	0.701	0.677	0.677	0.737	0.760	0.654	1.000	0.985	0.746	0.737
Specificity	0.390	0.369	0.500	0.369	0.323	0.323	0.419	0.478	0.307	0.000	0.000	0.455	0.458
PosPred Value	0.723	0.685	0.715	0.685	0.677	0.677	0.669	0.731	0.664	0.677	0.674	0.723	0.754
NegPred Value	0.371	0.387	0.597	0.387	0.323	0.323	0.500	0.516	0.297	NaN	0.000	0.484	0.436
Prevalence	0.693	0.662	0.615	0.662	0.677	0.677	0.615	0.651	0.677	0.677	0.677	0.656	0.693
Detection Rate	0.490	0.464	0.484	0.464	0.458	0.458	0.453	0.495	0.443	0.677	0.667	0.490	0.510
Detection Prevalence	0.677	0.677	0.677	0.677	0.677	0.677	0.677	0.677	0.667	1.000	0.990	0.677	0.677
Balanced Accuracy	0.548	0.535	0.644	0.535	0.500	0.500	0.578	0.619	0.480	0.500	0.492	0.600	0.597

the feature-class contributions when the learning model changes. This result in subset's to become ineffective for classification or the classification accuracy may get affected. Rank based methods are already introduced involving classification accuracy scores which are derived from distance and correlation values. The benefits of involving wrapper methods are its capability to identify the causative features, their interaction with other features. In our proposed work, using a non-linear method, the features and its effects are measured. These measurements can define the nature of a feature to be relevant or irrelevant to the classes available. The aggregated information maximization function describes the feature and its class contribution within a k subset. The main advantage of this function is to aggregate the feature relevancy to different classes available. While different information functions such as (MI) Mutual Information, (NMI) Normalized Mutual Information, (AMI) Adjusted Mutual Information, (SMI) Standardized Mutual Information, (ARI) Adjusted Rand Index are already introduced for feature selection in Romano et al. [24]. These functions tend to identify the relation between the feature and a class but they suffer from a lower discriminating power between the classes. To overcome this, our proposed method aggregates the feature relevancy to different classes and forms a subset. The performance accuracy of a classifier can be improved with our proposed feature selection method.

6 Conclusion

In this study, classification of heart disease is applied to sixteen algorithms using South African heart disease dataset and their performances are interpreted by plotting accuracy, recall and precision values. The classification of heart disease is done using nine features, and the role of classifiers in disease prediction is very important in treating heart disease. The accuracy of classification could improve the treatment quality and treatment process. Further identifying classification techniques using model

Table 4: Confusion Matrix for machine learning algorithms

CLUSTERING TECHNIQUES	CONFUSION MATRICES(PREDICTED)	
	TRAINING SET	TESTING SET
K-means	1 2 1 132 40 2 52 45	1 2 1 94 36 2 39 23
K-medoids	1 2 1 123 49 2 51 46	1 2 1 89 41 2 38 24
K-modes	1 2 1 107 65 2 70 27	1 2 1 93 37 2 25 37
PAM	1 2 1 123 49 2 51 46	1 2 1 89 41 2 38 24
CLARANS	1 2 1 66 106 2 58 39	1 2 1 88 42 2 42 20
CLARA	1 2 1 76 96 2 69 28	1 2 1 88 42 2 42 20
FCM	1 2 1 110 62 2 44 53	1 2 1 87 43 2 31 31
Ward's	1 2 1 169 3 2 93 4	1 2 195 35 2 30 32
ROCK	1 2 189 52 2 83 45	1 2 1 85 43 2 45 19
DBSCAN	1 2 1 172 97 2 0 0	1 2 1 130 62 2 0 0
OPTICS	1 2 1 172 97 2 0 0	1 2 1 128 62 2 2 0
EM	1 2 1 93 79 2 86 11	1 2 194 36 2 32 30

based methods can help unveil hidden patterns and information and help medical experts to diagnose disease earlier. Further, the importance of utilizing recent data in classifying demands more of an efficient algorithm and thus we extend our work to develop a novel hybrid algorithm and study its performances across other algorithms and datasets with good accuracy percentage.

Table 5: Performance Measures

MEASURES	K-means	K-medoids	K-modes	PAM	CLARANS	CLARA	FCM	
Precision	0.71	0.70	0.79	0.70	0.68	0.68	0.74	
Recall	0.39	0.37	0.50	0.37	0.32	0.32	0.42	
Accuracy	0.55	0.54	0.64	0.54	0.50	0.50	0.58	
MEASURES	WARD'S	ROCK	DBSCAN	OPTICS	EM	SVM	Decision tree	K-nn
Precision	0.76	0.65	1.00	0.98	0.75	0.45	0.91	0.74
Recall	0.48	0.31	0.00	0.00	0.45	0.42	0.56	0.46
Accuracy	0.62	0.48	0.50	0.49	0.60	0.44	0.73	0.60

Acknowledgement

The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] Boros E, Hammer PL, Ibaraki T, Kogan A, Mayoraz E, Muchnik I. An implementation of logical analysis of data.IEEE Transactions on Knowledge and Data Engineering. 2000 Mar;12(2):292-306.(2000): 292-306.

[2] Das R, Turkoglu I, Sengur A. Effective diagnosis of heart disease through neural networks ensembles. Expert systems with applications. 2009 May 31;36(4):7675-80.

[3] Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid JJ, Sandhu S, Guppy KH, Lee S, Froelicher V. International application of a new probability algorithm for the diagnosis of coronary artery disease. The American journal of cardiology. 1989 Aug 1;64(5):304-10.

[4] El-Hanjouri M, Alkhaldi W, Hamdy N, Alim OA. Heart diseases diagnosis using HMM. In Electrotechnical Conference, 2002.MELECON 2002. 11th Mediterranean 2002 (pp. 489-492). IEEE.

[5] Skalak DB. Prototype selection for composite nearest neighbor classifiers (Doctoral dissertation, University of Massachusetts at Amherst), 1997.

[6] Avci E. A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier.Expert Systems with Applications. 2009 Sep 30;36(7):10618-26.

[7] Doyle OM, Temko A, Marnane W, Lightbody G, Boylan GB. Heart rate based automatic seizure detection in the newborn. Medical engineering & physics. 2010 Oct 31;32(8):829-39.

[8] Gamboa AL, Mendoza MG, Orozco RE, Vargas JM, Gress NH. Hybrid fuzzy-SV clustering for heart disease identification. In Computational Intelligence for Modelling, Control and Automation, 2006 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on 2006 Nov 28 (pp. 121-121).IEEE.

[9] Maglogiannis I, Loukis E, Zafiropoulos E, Stasis A. Support vectors machine-based identification of heart valve diseases using heart sounds. Computer methods and programs in biomedicine. 2009 Jul 31;95(1):47-61.

[10] Obayya M, Abou-Chadi F. Data fusion for heart diseases classification using multi-layer feed forward neural network. In Computer Engineering & Systems, 2008.ICCES 2008. International Conference on 2008 Nov 25 (pp. 67-70). IEEE.

[11] Zheng J, Jiang Y, Yan H. Committee machines with ensembles of multilayer perceptron for the support of diagnosis of heart diseases. In Communications, Circuits and Systems Proceedings, 2006 International Conference on 2006 Jun 25 (Vol. 3, pp. 2046-2050).IEEE.

[12] Kim BH, Lee SH, Cho DU, Oh SY. A proposal of heart diseases diagnosis method using analysis of face color.InAdvanced Language Processing and Web Information Technology, 2008.ALPIT'08. International Conference on 2008 Jul 23 (pp. 220-225). IEEE.

[13] Tripoliti EE, Papadopoulos TG, Karanasiou GS, Naka KK, Fotiadis DI. Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques. Computational and structural biotechnology journal.2017 Dec 31;15:26-47.

[14] Jovic A, Bogunovic N. Electrocardiogram analysis using a combination of statistical, geometric, and nonlinear heart rate variability features. Artificial intelligence in medicine. 2011 Mar 31;51(3):175-86.

[15] Melillo P, Fusco R, Sansone M, Bracale M, Pecchia L. Discrimination power of long-term heart rate variability measures for chronic heart failure detection. Medical & biological engineering & computing. 2011 Jan 1;49(1):67-74.

[16] Yu SN, Lee MY. Conditional mutual information-based feature selection for congestive heart failure recognition using heart rate variability.Computer methods and programs in biomedicine. 2012 Oct 31;108(1):299-309.

[17] Liu G, Wang L, Wang Q, Zhou G, Wang Y, Jiang Q. A new approach to detect congestive heart failure using short-term heart rate variability measures.PloS one. 2014 Apr 18;9(4):e93399.

[18] Narin A, Isler Y, Ozer M. Investigating the performance improvement of HRV Indices in CHF using feature selection methods based on backward elimination and statistical significance. Computers in biology and medicine.2014 Feb 1;45:72-9.

[19] Zheng Y, Guo X, Qin J, Xiao S. Computer-assisted diagnosis for chronic heart failure by the analysis of their cardiac reserve and heart sound characteristics. Computer methods and programs in biomedicine. 2015 Dec 31;122(3):372-83.

[20] Masetic Z, Subasi A. Congestive heart failure detection using random forest classifier. Computer methods and programs in biomedicine.2016 Jul 31;130:54-64.

[21] Bohacik J, Kambhampati C, Davis DN, Cleland JG. Alternating decision tree applied to risk assessment of heart failure patients. Journal of Information Technologies. 2013;6(2):25-33.

[22] Fahad A, Alshatri N, Tari Z, Alamri A, Khalil I, Zomaya AY, Foufou S, Bouras A. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE transactions on emerging topics in computing. 2014 Sep;2(3):267-79.
[23] Rossouw JE, Du Plessis JP, Benad AJ, Jordaan PC, Kotze JP, Jooste PL, Ferreira JJ. Coronary risk factor screening in three rural communities.The CORIS baseline study. South African medical journal.1983 Sep; 64(12):430-6.
[24] Romano S, Bailey J, Nguyen V, Verspoor K. Standardized mutual information for clustering comparisons: one step further in adjustment for chance. International Conference on Machine Learning.2014 Jan 27: 1143-1151.

M. Chandralekha is currently pursuing her Ph.D. in the Department of Computer Applications, University College of Engineering, Bharathidasan Institute of Technology Campus, Anna University,Tiruchirappalli, Tamilnadu,India. She obtained her Bachelor's degree from Amrita Vishwa Vidyapeetham, Amrita School of Engineering, Ettimadai, Coimbatore, Tamilnadu, India in 2014. She received her Master's degree from PSG College of Technology, Coimbatore, Tamilnadu, India in 2016. She is currently doing research in the area of Distributed Computing. Her research areas include Data mining, Mobile Computing,Grid and Cloud Computing, Information retrieval and Open source systems. She has published many research articles in reputed international journals.

N. Shenbagavadivu is an Assistant Professor in the Department of Computer Applications, University College of Engineering, Bharathidasan Institute of Technology Campus, Anna University, Tiruchirappalli, Tamilnadu, India. She received her Ph.D. in the year of 2011. Her research areas include Distributed Computing, Web services, Mobile and Cloud Computing. She has published more than 50 papers in international, national journals and conferences. She has been a reviewer for many national and international journals.