Nat. Volatiles & Essent. Oils, 2021; 8(5): 2136 - 2142

NVEO
Natural Volatiles & Essential Oils

Early Prognosis of Coronary Heart Disease using Ensemble Classifiers: A Comparative Analysis

H.S.Niranjana Murthy , M.N.Manjunatha, UmesharaddyRadder

1Dept. Of ElectronicsIns trumentation Engineering Institute of Technology Ramaiah Institute of Technology Bangalore-54, IndiaBangalore-54, India hasnimurthy@msrit.edu

2 Dept.of Chemistry Ramaiah Institute of Technology Bangalore-54, India
mnmanjunathmsrit@gmail.com
3Dept. of Electronics & Telecommunication Engineering Ramaiah Institute of Bangalore-54,India
umesh.reddy@msrit.edu

Abstract:

Machine Learning techniques are extensively used in health care especially for disease prediction. This paper presents a comparison of performance of various Ensemble classifiers for early detection of Coronary Heart Disease (CHD) based on risk factors. This paper focuses on the Bagging, Boosting and Subspace Ensemble Classifiers for detecting CHD. The performance of these ensemble classifiers is compared with performance indices such as accuracy, precision, recall and F1 score. K-Fold's validation is adopted to randomize the data and to obtain the consistency of results. In the current research work, the experimentation has been carried out on the datasets acquired from UCI dataset. From the experimentation results, it is observed that Bagged Trees Ensemble classifier provides a highest classification accuracy, precision, recall and F1 score of 95.5 %, 0.95, 0.97 and 0.95 respectively for identifying CHD. The result also depicted that the Bagged Trees Ensemble classifier outperformed in comparison with the traditional classifiers. The current work is useful for physicians to detect the coronary heart disease at early stages.

Keywords—*Coronary heart disease, Ensemble classifier, Bootstrap Aggregator, Ada-boost, Subspace discriminant.*

I. INTRODUCTION

The primary cause for cardiac ischemia is the obstruction of coronary artery leading to coronary heart disease which ultimately results in heart stroke or heart attack. Chest pains occur as an initial symptom, when the blood received by the heart muscles is insufficient. The risk level of coronary heart disease can be estimated from the multivariate clinical risk factors. These factors include cholesterol, LDL-C, HDL-C, blood pressure, and diabetes. The other factors which contribute to coronary artery disease are family history of premature coronary heart disease, obesity, left ventricular hypertrophy, and estrogen replacement therapy (ERT). Medically, it is established that several years of follow-up of these risk factors have facilitated the diagnosis of CHD.

Although, the ECG and CT scans help in the detection of CHD, the high cost and infeasibility of these techniques have resulted in the 17 million deaths of patients every year due to CHD [1]. The Lancet study on global burden of disease study indicated CHD as the chronic disease [2]. The risk factors such as blood pressure, smoking, alcohol intake is generally found in developed countries where 85% of deaths occur due to chronic diseases [3]. Also, the chronic diseases are increasing in the developing countries due to unhealthy diet habits, sedentary life styles and malnutrition [4]. Even though, the angiography is the conventional technique available for detecting CHD, but it requires strong technical knowledge and high cost [5]. To mitigate these issues, the machine learning algorithms can be developed for detecting the CHD by using the risk factors [6].

Generally, the various classifier systems or the ensemble-based strategies are increasingly alluring thought as they lessen the reduced determination plausibility [7]. The ensemble joins a lot of classifiers that may create unrivalled grouping execution contrasted with single classifier. The gathering of

classifiers is considered based on (i) classifier choice, where the classifier with the superior execution is chosen as the last yield, or (ii) classifier combination, where the yields of the separate classifiers are consolidated to decide an on classifiers [8]. The most common mix rules incorporate the weighted majority voting, Borda tally [9]. The determination of the gathering size includes a harmony between the precision and speed, where over-trained sorting may happen with too huge gatherings and bigger ensembles set aside lengthy time for forecast.

However, as indicated by the past examinations, not many computerized classifiers using ensembles for early detection of CHD have been accounted for. Likewise, the precision of classifiers for foreseeing CHD was constrained to around 80%. Moreover, Ensemble based strategy has not been incorporated for recognizing heart disease. Thus, the present work targets improving the forecast exactness by applying Ensemble classifiers on the UCI dataset for detecting CHD. The proposed ensemble classifier utilized the extricated statistical highlights. In addition, a near investigation of various ensembles, to be specific the bagged, boosted and the subspace ensemble classifiers, is additionally included.

The structure of the rest of the segments is as per the following. Segment 2 highlights the technique and proposed strategy for the current work. Area 3 exhibits the acquired outcomes with near examinations. At long last, Section 4 concludes the proposed investigation.

II. METHODOLOGY

A. Heart Disease Dataset

The estimation of threat of cardiac ischemia is carried by evaluation of level of risk of coronary artery disease by using multivariate clinical risk factor features obtained from Cleveland database [10]. The dataset used in this study consists of clinical data of subjects having the symptoms of coronary heart disease and consists of four databases: Cleveland, Hungary and Switzerland. It contains 76 attributes, but all earlier works refer to using a subset of 14 of them. The target field refers to the presence of heart disease in the patient. In the first stage, analysis of data is carried out by discarding missing values, wrong type values & outliers. Table 1 depicts the characteristics of clinical risk factors of heart disease database.

TABLE 1 ATTRIBUTES OF CLINICAL RISK FACTORS OF CORONARY HEART DISEASE DATABASE

Sl. No.	Parameters
1.	Age in years
2.	Sex (1 = male; 0 = female)
3.	Chest pain types: 1= typical angina 2=atypical angina; 3=non-angina pain; 4=asymptomatic
4.	Resting B.P in mm of Hg
5.	Serum cholesterol in mg/dl
6.	Fasting blood sugar > 120mg/dl 1 = True; 0 = False
7.	Resting ECG results 0 = Normal; 1 = T wave inversion/ST elevation; 2 = left ventricular Hypertrophy
8.	Maximum heart rate achieved
9.	Exercise induced angina 1 = yes; 0 = no
10.	ST depression induced by exercise
11.	Slope of peak exercise ST segment 1 = up sloping; 2 = flat; 3= down sloping
12.	Number of major vessels (0 – 3)
13.	3= Normal; 6 = fixed defect; 7 = reversible defect
14.	diagnosis of heart disease (angiographic disease status) Value 0: No disease Value 1: Mild Value 2: Moderate Value 3: Severe

B. Proposed Approach

The main aim of this paper is to improve the accuracy of predicting the coronary heart disease by using various ensemble techniques

The information acquired by UCI database is pre-handled to expel unessential and missing information. Further, the prevalent highlights are tried on ensemble classifier for different execution measures. Steps engaged with the proposed algorithm are as per the following.

Stage 1: Load the UCI information.

Stage 2: Pre-handling of information to expel superfluous and anomalies.

Stage 3: Apply feature choice measure on data.

Stage 4: Remove least positioned characteristics and keep dominating features.

Stage 5: Apply ensemble classifiers on prevalent highlights (Bagging, Boosting, and Subspace)

Stage 6: Measure the presentation of proposed technique.

Ensemble learning joins a few models for improving the expectation execution, which has a few methodologies such as (i) Bagging, which makes a lot of models that prepared on an irregular data, then the forecasts are accumulated/consolidated for conclusive expectation utilizing voting, and (ii) boosting depends on voting of different models, where it assess the developed models dependent on their exhibition. (iii) subspace technique, which randomizes the learning calculation by choosing a subset of highlights haphazardly before training, and afterward the models' yields are consolidated by majority vote Bagged Tree Ensemble Classifier

C. Bagging Ensemble Classifier

The bagging strategy is valuable for both regression and factual characterization. Bagging is utilized with choice trees, where it fundamentally raises the solidness of models in the decrease of difference and improving precision, which takes out the test of overfitting. Bagging in group AI takes a few frail models, collecting the expectations to choose the best forecast. The feeble models have some expertise in particular segments of the component space, which empowers bagging influence expectations to originate from each model to arrive at the most extreme reason.

Bagging is made out of two sections: aggregation and bootstrapping [11]. Bootstrapping is a sampling technique, where an example is picked out of a set, utilizing the replacement strategy. The learning calculation is then run on chosen samples. The bootstrapping method utilizes sampling with replacements to make the choice methodology totally arbitrary. At the point when an example is chosen without replacements, the ensuing choices of factors are constantly subject to the past choices, henceforth making the models non-arbitrary.

Model forecasts experience aggregation to consolidate them for the last expectation to consider from all the results conceivable. The aggregation should be possible dependent on the all-out number of results or on the likelihood of expectations got from the bootstrapping of each model in the system.

The outline of Bootstrapping bagging is shown in Algorithm 1. Given the parameter m, which is the number of learner models, the algorithm provides the bootstrap samples from the heart disease data D to the m learners by means of row sampling with replacement. The learning algorithm L is applied to produce the output from the multiple classifier models Bi, which are aggregated to create the final Classifier B

Algorithm 1: Bootstrap Aggregator Algorithm

Input:
Training set of heart disease data D
Number of learner model m
Learning algorithm L
Procedure:
for i=1 →m do
D_i ←Bootstrap sample from D
Component Classifier B_i L(D_i)
end for
Output:
Final Classifier B by majority voting

Bagging offers the merit of permitting numerous weak classifier to join endeavours to exceed a single strong classifier. It likewise helps in the decrease of difference, consequently disposing of the overfitting of models in the strategy. One impediment of bagging is that it presents lost interpretability of a model. The resultant model can encounter more inclination when the best possible technique is disregarded. In spite of bagging being exceptionally precise, it very well may be computationally costly and this may debilitate its utilization in specific cases.

D. Boosting Ensemble Classifier

Boosting is a calculation that helps in decreasing difference and inclination in an AI gathering [12]. The calculation helps in the transformation of weak learners into strong learners by joining N number of learners. Boosting likewise can improve model expectations for learning calculations. The weak learners are consecutively revised by their antecedents and, all the while, they are changed over into powerful learners. Boosting can take a few structures, for example, i) Adaboost, ii) Gradient Boosting and iii) RUS(Random under sampling) Boost.

Adaboost targets consolidating a few frail learners to frame a solitary solid learner. Adaboost focuses on weak learners, which are frequently decision trees with just one split and are normally alluded to as decision stumps. The principal decision stump in Adaboost contains perceptions that are weighted similarly. A mistake saw in past models is balanced with weighting until an exact indicator is made.

Gradient boosting successively adds predictors to the ensemble and follows the grouping in adjusting going before predictor to show up at a precise indicator toward the finish of the technique. Slope boosting uses the gradient descent procedure in the learner's expectations. The past blunder is featured, and, by consolidating one weak learner to the following learner, the mistake is decreased essentially after some time.

RUS Boost is a calculation to deal with class irregularity issue in information with discrete class marks.

It utilizes a blend of RUS and the standard boosting system AdaBoost, to all the more likely model the minority class by eliminating majority class samples. Boosting models are being fundamentally engaged at decreasing inclination, the base models that are frequently considered for boosting are models with low fluctuation yet high predisposition.

E. Subspace Ensemble Classifier

Subspace calculation utilizes either k-nearest neighbour learner or discriminant analysis classifier as base classifier. Subspace model makes an outfit of discriminant classifiers utilizing random subspace calculation. It is useful for some predictors, low on memory usage, moderately quick for fitting and forecast yet the exactness fluctuates relying on the data. The subspace ensemble proposed by Ho [13], uses random choice of feature subspaces to develop singular classifiers. This strategy can exploit high dimensionality, and is a compelling countermeasure for the conventional issue of the scourge of dimensionality. This strategy results in the high ensemble assortment, which make up the reduction of precisions in individual classifiers [14]. In random subspace, subspaces are selected arbitrarily from the primary feature space, utilizing as first training set. The KNN classifier is a customary classification rule, which assigns the label of a test sample with the majority label of the training set [15]. As the objective of this article is to assess the working of ensembles, there is no requirement for us to extravagantly tune k.

III Results and Discussions

Table II depicts the percentage accuracy, precision, recall and F1 score of Ensemble classifiers for the Cleveland heart disease datasets obtained with the technique as discussed in previous section. From the table, it is seen that the Bagged trees ensemble classifier has outperformed and provides a highest performance metric corresponding to 95.5 %, 0.95, 0.97 and 0.95 of accuracy, precision, recall and F1 score.

TABLE II
PERFORMANCE METRIC OF ENSEMBLE CLASSIFIERS WITH HEART DISEASE DATASETS

Ensemble Method	**Training Accuracy (%)**	**Testing Accuracy (%)**	**Precision**	**Recall**	**F1 Score**
AdaBoost	83	86.7	1	0.9	0.95
Bagged Trees	100	95.5	0.95	0.97	0.95
Subspace Discriminant	64.7	63.6	1	0.8	0.89
RUS Boosted Trees	78.5	81.8	0.88	0.96	0.92

The experimental results exhibited that Bagging trees ensemble classifier performs the best as compared to other ensemble classifiers on all types of datasets. Adaboost displayed the second-best performance and the subspace discriminant classifier displayed worst performance on Cleveland heart disease dataset.

The confusion matrix of bagged tree ensemble classifier on heart disease dataset is illustrated in Fig. 2. Also, the ROC curve is demonstrated in Fig. 3, which shows a highest value of AUC as 99% for the Bagged tree ensemble classifier on heart disease dataset. Fig.4 shows the comparison of accuracy of ensemble classifiers on Cleveland, Hungarian and Switzerland heart disease datasets. The comparison

indicates that the Bagged tree ensemble classifier performs on all types of datasets with highest accuracy.

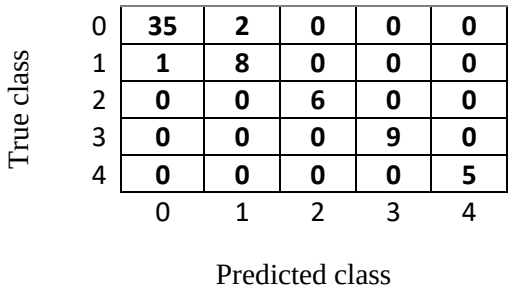

Fig. 2 Confusion Matrix for Bagged Trees Ensemble Classier on Heart disease Dataset.

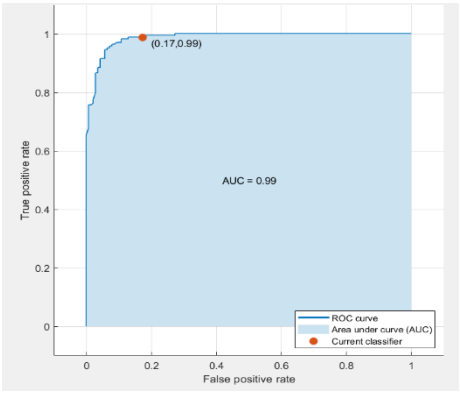

Fig. 3 ROC curve for Bagged Trees Ensemble Classifier on Heart disease Dataset.

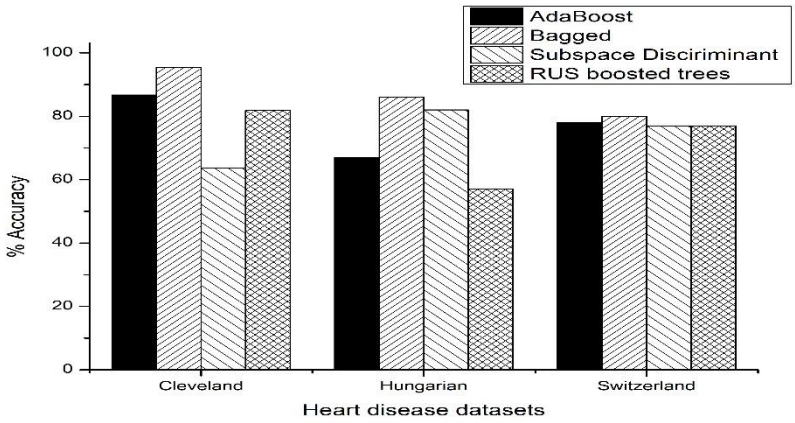

Fig. 4 Accuracy measure of Ensemble classifiers on various Datasets

IV. CONCLUSION

This paper exhibited a comparative analysis of Ensemble Classifier models for identifying the coronary heart disease. The feature vector is pre-processed for removing outliers and feature selection is carried. Various architectures of Bagging, Boosting and Subspace Ensemble Classifiers with different numbers of splits and number of learners are considered for identifying heart disease with highest accuracy. The results clearly confirmed that the Bagged Trees Ensemble Classifier model with 20 splits and 30 number of learners has outperformed with 95.5 %, 0.95, 0.97 and 0.95 of accuracy, precision, recall and F1 score respectively, using Cleveland heart disease data set. Also, the Bagged trees classifier model outperformed on Hungarian and Switzerland heart disease data sets. Hence, the Bagged trees ensemble classifier model can be used for early prognosis of heart disease with high accuracy.

REFERENCES

[1] Judith Mackay, George Mensah, Shanthi Mendis and Kurt Greenland, "The atlas of heart disease and stroke", World Health Organization, 2004.
[2] Huse, O., Hettiarachchi, J., Gearon, E., Nichols, M., Allender, S., & Peeters, A., "Obesity in Australia", *Obesity Research & Clinical Practice*, *12*(1), pp. 29-39, 2018
[3] Schmidt H, "Chronic Disease Prevention and Health Promotion", Public Health Ethics: Cases, Springer, chapter 5, PMID, 2016.
[4] *Working together for Health*, the World Health Report, World Health Organisation, pp. 19, 2006.
[5] Gonsalves, A. H., Thabtah, F., Mohammad, R. M., & Singh, G., "Prediction of Coronary Heart Disease using Machine Learning", *Proceedings of the International Conference on Deep Learning Technologies, 2019*.
[6] Patil SB, Kumaraswamy Y, "Intelligent and effective heart attack prediction system using data mining and artificial neural network", Eur. J. Sci. Res., 31, pp. 642–656, 2009.
[7] Kuncheva Li, Combining pattern classifiers: methods and algorithms, Newyork: Wiley, 2004.
[8] Woods K, Kegelmeyer WP, Bowyer K.,"Combination of multiple classifiers using local accuracy estimates". IEEE Trans Pattern Anal Mach Intell., vol.19, No.4, pp.405–410, 1997.
[9] Polikar R. Ensemble based systems in decision making. IEEE Circuits Syst. Mag., vol.6, No.3, pp.21–45, 2006.
[10] Dua, D., and C. Graff, "UCI Machine Learning Repository." Irvine, CA: University of California, School of Information and Computer Science, 2019.
[11] Efron B, Tibshirani R, An Introduction to the Bootstrap. Chapman & Hall, New York, 1993
[12] Dietterich TG, "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization", Machine Learning, Vol. 40, No.2, pp.139–157, 2000.
[13] Ho TK, "The Random subspace method for constructing decision forests", IEEE Trans Pattern Analysis and Machine Intelligence 1998; vol. 20, No.8, pp.832–844, 1998.
[14] Tsymbal, A., Pechenizkiy, M., Cunningham, P., "Diversity in search strategies for ensemble feature selection", Information Fusion, vol 6, No. 1, pp.83–98, 2005.
[15] Duda, R.O., Hart, P.E., Stork, D.G., Pattern Classification, 2nd ed. John Wiley and Sons, New York, 2000.