Third Year Seminar Report

Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques

by

Sharvil Shailesh Bakshi
(T.E. Roll No.: 08 Year: (2022-2023))

Guide

Prof. Rupali Bora

Department of Information Technology
K.K.Wagh Institute of Engineering Education and Research
Nashik -422003

Abstract

Heart disease is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. In the study it is also seen ML techniques being used in recent developments in different areas of Machine Learning (ML). Various studies give only a glimpse into predicting heart disease with ML techniques. In this paper, a novel method is proposed that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. An enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM)is produced.[1]

Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction

It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Various techniques in data mining and neural networks have been employed to find out the severity of heart disease among humans. The severity of the disease is classified based on various methods like K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naive Bayes (NB). The nature of heart disease is complex and hence, the disease must be handled carefully. Not doing so may affect the heart or cause premature death. The perspective of medical science and data mining are used for discovering various sorts of metabolic syndromes. Data mining with classification plays a significant role in the prediction of heart disease and data investigation.
Various methods have been used for knowledge abstraction by using known methods of data mining for prediction of heart disease. In this work, numerous readings have been carried out to produce a prediction model using not only distinct techniques but also by relating two or more techniques. These amalgamated new techniques are commonly known as hybrid methods. Neural networks using heart rate time series is introduced. This method uses various clinical records for prediction such as Left bundle branch block (LBBB), Right bundle branch block (RBBB), Atrial fibrillation (AFIB), Normal Sinus Rhythm (NSR), Sinus bradycardia (SBR), Atrial flutter (AFL), Premature Ventricular Contraction (PVC)), and Second degree block (BII) to find out the exact condition of the patient in relation to heart disease. The rest of the paper is organized as follows, Section II discusses heart related works, existing methods and techniques available also HRFLM Data pre-processing followed by feature selection, classification modeling and performance measure are discussed. Section III gives the algorithms used and the experimental setup. Overview of results are plotted in Section IV in the Conclusion.[1]

1.2 Motivation

Well starting with searching research papers on internet, finally found the IEEE website with a huge domain of different papers with different authors and researchers all over the world. Now the main challenge was to select a paper with suitable technical work and also a unique concept out of the box.
Also one motivation was the interest in the machine learning domain which further filtered my searches in the right direction. Machine learning domain has vast scope in the field of medical science.
In the ocean of millions international authors finally found my way searching our Indian authors for the paper. And this was the driving force behind selecting the paper at one read itself.

1.3 Aim and Objectives

Various studies give only a glimpse into predicting heart disease with machine learning techniques. In this paper, a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease is produced.
Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values. Predictive modelling is a mathematical process used to predict future events or outcomes by analysing patterns in a given set of input data. It is a crucial component of predictive analytics, a type of data analytics which uses current and historical data to forecast activity, behavior and trends.[2]

1.4 Seminar Topic

Neural networks are generally regarded as the best tool for prediction of diseases like heart disease and brain disease. The proposed method which is used has 13 attributes for heart disease prediction. The results show an enhanced level of performance compared to the existing methods in works like. The Carotid Artery Stenting (CAS) has also become a prevalent treatment mode in the medical field during these recent years. The CAS prompts the occurrence of major adverse cardiovascular events (MACE) of heart disease patients that are elderly. Their evaluation becomes very important. Results are generated using a Artificial Neural Network (ANN), which produces good performance in the prediction of heart disease. Neural network methods are introduced, which combine not only posterior probabilities but also predicted values from multiple predecessor techniques. This model achieves an accuracy level of up to 89.01% which is a strong results compared to previous works. seminars.

In this work, we introduce a technique we call the Hybrid Random Forest with Linear Model (HRFLM). The main objective of this research is to improve the performance accuracy of heart disease prediction. Many studies have been conducted that results in restrictions of feature selection for algorithmic use. In contrast, the HRFLM method uses all features without any restrictions of feature selection. Here we conduct experiments used to identify the features of a machine learning algorithm with a hybrid method. The experiment results show that our proposed hybrid method has stronger capability to predict heart disease compared to existing methods. So the topic name goes EFFECTIVE HEART DISEASE PREDICTION USING HYBRID MACHINE LEARNING TECHNIQUES.[1]

1.5 Related Work

There is ample of related work in the fields directly related to this paper. ANN has been introduced to produce the highest accuracy prediction in the medical field. The back propagation multilayer perception (MLP) of ANN is used to predict heart disease. The obtained results are compared with the results of existing models within the same domain and found to be improved. The data of heart disease patients collected from the UCI laboratory is used to discover patterns with NN, DT, Support Vector machines SVM, and Naive Bayes. The results are compared for performance and accuracy with these algorithms. The proposed hybrid method returns results of 86.8% for F-measure, competing with the other existing methods. The new approaches presented here decrease the cost and improve the prediction of heart disease in an easy and effective way. The various different research techniques considered in this work for prediction and classification of heart disease using ML and deep learning (DL) techniques are highly accurate in establishing the efficiency of these methods.[1]

Chapter 2

Methodology

2.1 Data Pre-Processing

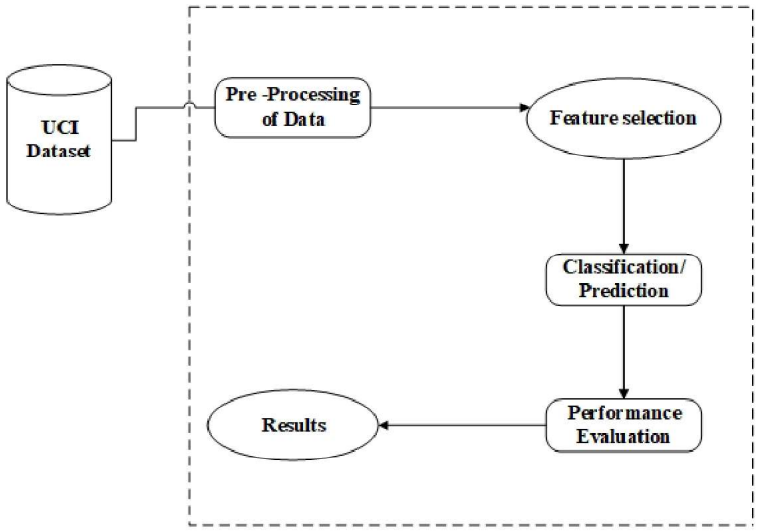

FIGURE 1. Experiment workflow with UCI dataset.

Figure 2.1: Workflow with UCI

Heart disease data is pre-processed after collection of various records. The dataset contains a total of 303 patient records, where 6 records are with some missing values. Those 6 records have been removed from the dataset and the remaining 297 patient records are used in preprocessing. The multiclass variable and binary classification are introduced for the attributes of the given dataset. The multi-class variable is used to check the presence or absence of heart disease. In the instance of the patient having heart disease, the figure, shows UCI dataset attributes detailed information.[1]

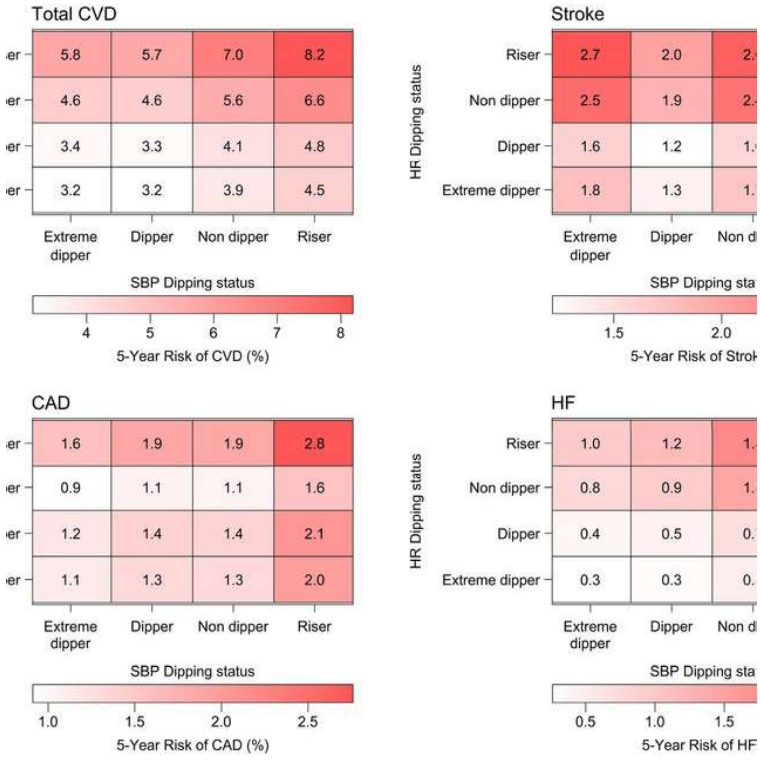

Figure 2.2: Chart

The pre-processing of data is carried out by converting medical records into diagnosis values. The results of data pre-processing for 297 patient records indicate that 137 records show the value of 1 establishing the presence of heart disease while the remaining 160 reflected the value of 0 indicating the absence of heart disease.[3]

2.2 Feature Selection and Reduction

From among the 13 attributes of the data set, two attributes pertaining to age and sex are used to identify the personal information of the patient. The remaining 11 attributes are considered important as they contain vital clinical records. Clinical records are vital to diagnosis and learning the severity of heart disease. As previously mentioned in this experiment, several (ML) techniques are used namely, NB, GLM, LR, DL, DT, RF, GBT.[1]

AGE	Numeric [29 to 77;unique=41;mean=54.4;median=56]
SEX	Numeric [0 to 1;unique=2;mean=0.68;median=1]
CP	Numeric [1 to 4;unique=4;mean=3.16;median=3]
TESTBPS	Numeric [94 to 200;unique=50;mean=131.69;median=130]
CHOL	Numeric [126 to 564;unique=152;mean=246.69;median=241]
FBS	Numeric [0 to 1;unique=2;mean=0.15;median=0]
RESTECG	Numeric [0 to 2;unique=3;mean=0.99;median=1]
THALACH	Numeric [71 to 202;unique=91;mean=149.61;median=153]
EXANG	Numeric [0 to 1;unique=2;mean=0.33;median=0.00]
OLPEAK	Numeric [0 to 6.20;unique=40;mean=1.04;median=0.80]
SLOPE	Numeric [1 to 3;unique=3;mean=1.60;median=2]
CA	Categorical [5 levels]
THAL	Categorical [4 levels]
TARGET	Numeric [0.00 to 4.00;unique=5;mean=0.94;median=0.00]

Figure 2.3: Feature Selection

2.3 Classification Modelling

The clustering of datasets is done on the basis of the variables and criteria of Decision Tree (DT) features. Then, the classifiers are applied to each clustered dataset in order to estimate its performance. The best performing models are identified from the above results based on their low rate of error. The performance is further optimized by choosing the DT cluster with a high rate of error and extraction of its corresponding classifier features. The performance of the classifier is evaluated for error optimization on this data set.[1]

1) DECISION TREES
Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

2) LANGUAGE MODEL
A language model is a probability distribution over sequences of words.

3) SUPPORT VECTOR MACHINE
Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms,

which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

4) RANDOM FOREST
Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.

5) NAIVE BAYES
Naive Bayes classifiers are a collection of classification algorithms based on Bayes Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

6) NEURAL NETWORKS
A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates.

7) K-NEAREST NEIGHBOUR
The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.[2]

2.4 Proposed Method HRFLM

In this study, an R studio rattle to perform heart disease classification of the Cleveland UCI repository is used. It provides an easy-to-use visual representation of the dataset, working environment and building the predictive analytics. ML process starts from a preprocessing data phase followed by feature selection based on DT entropy, classification of modeling performance evaluation, and the results with improved accuracy. The feature selection and modeling keep on repeating for various combinations of attributes. Graph shows the UCI dataset detailed information with attributes used. The performance of each model generated based on 13 features and ML techniques used for each iteration of the results and performance are recorded.[3]

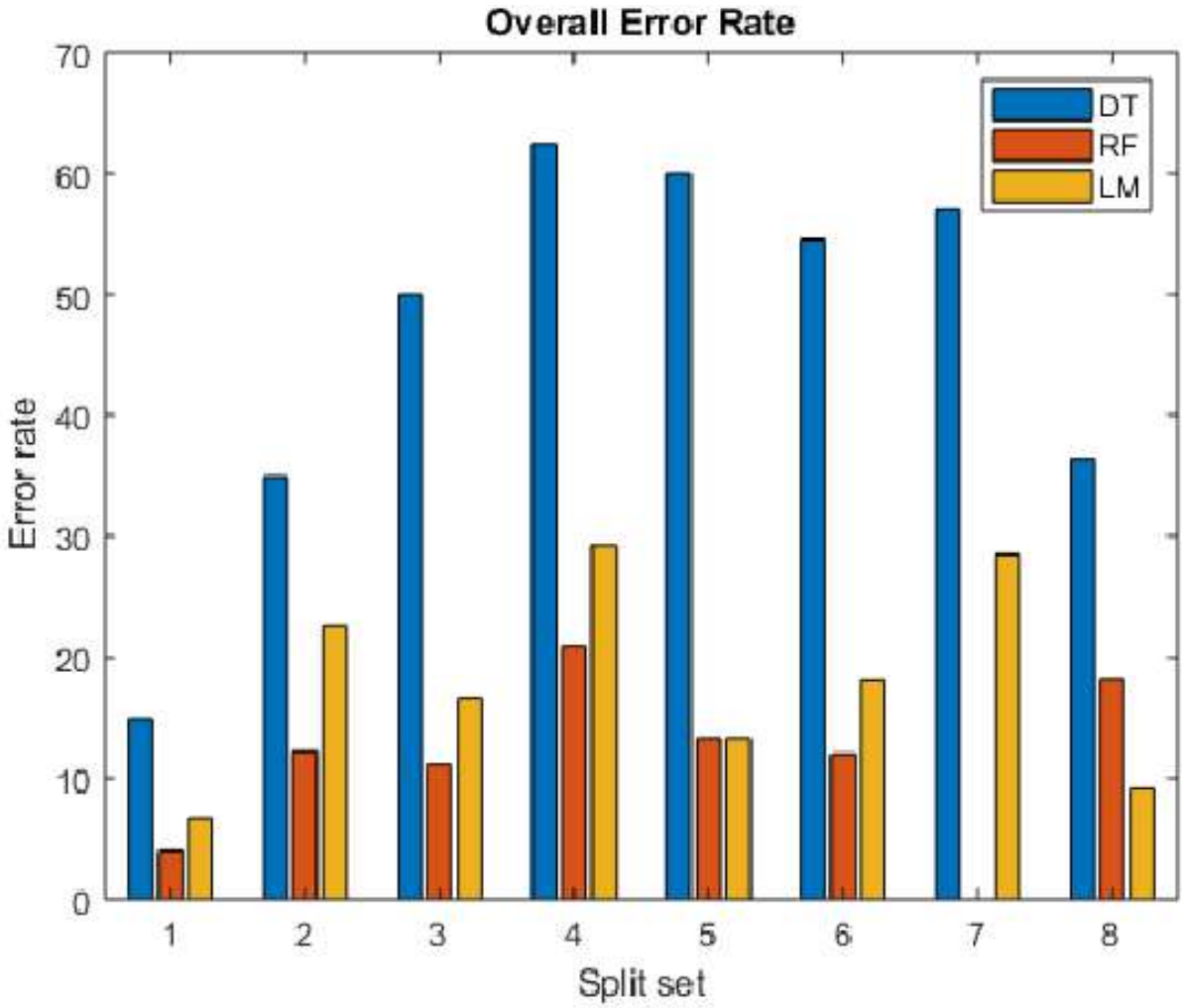

Figure 2.4: Error rate of HRFLM model

Chapter 3

Algorithms and Results

3.1 Algorithm 1 : Decision Tree-Based Partition

Require: Input: D dataset features with a target class
for features do
for Each sample do
Execute the Decision Tree algorithm
end for
Identify the feature space f1, f2, . . . , fx of dataset UCI.
end for
Obtain the total number of leaf nodes l1, l2, l3,. . . , ln with its constraints
Split the dataset D into d1, d2, d3,. . . , dn based on the leaf nodes constraints.
Output: Partition datasets d1, d2, d3,. . . , dn[1]

3.2 Algorithm 2 : Apply ML to Find Less Error Rate

Require: Input: Datasets with partition d1, d2, d3,. . . , dn
for apply the rules do
On the dataset R(d1,d2,d3,. . . . . dn)
end for
Classify the dataset based on the rules C(R(d1),R(d2) . . . . . .R(dn))
Output: Classified datasets with rules C(R(d1),R(d2) . . . . . .R(dn))[1]

3.3 Algorithm 3 : Feature Extraction Using Less Error Classifier

Require: Input: Classified datasets
C(R(d1),R(d2). . . . . .R(dn))

for Find out min error rate from the input do
Min(C(R(d1),R(d2). . . . . .R(dn)))
end for
Find out max(min) error rate from the classifier.
Output: Features with classified attributes F(d1, d2, d3, . . . , dn)[1]

3.4 Evaluation Results

Models	Accuracy	Classification error	Precision	F-measure	Sensitivity	Specificity
Naive Bayes	75.8	24.2	90.5	84.5	79.8	60.0
Generalized Linear Model	85.1	14.9	88.8	91.6	94.9	20.0
Logistic Regression	82.9	17.1	89.6	90.2	91.1	25.0
Deep Learning	87.4	12.6	90.7	92.6	95	33.3
Decision Tree	85	15.0	86	91.8	98.8	0.0
Random Forest	86.1	13.9	87.1	92.4	98.8	10.0
Gradient Boosted Trees	78.3	21.7	94.1	86.8	80.7	60.0
Support Vector Machine	86.1	13.9	86.1	92.5	100	0.0
VOTE	87.41	12.59	90.2	84.4	-	-
HRFLM (proposed)	**88.4**	**11.6**	**90.1**	**90**	**92.8**	**82.6**

Figure 3.1: Results

In HRFLM, a computational approach with the three association rules of mining namely, apriori, predictive and Tertius to find the factors of heart disease on the UCI Cleveland dataset is used. The available information points to the deduction that females have less of a chance for heart disease compared to males. In heart diseases, accurate diagnosis is primary. But, the traditional approaches are inadequate for accurate prediction and diagnosis. In the UCI data set 297 instances of patient records, in total, are considered of which 252 records are used for training and the remaining for testing. The results have been located to be satisfying based on the assessment. Heart disease prediction with SVM and ANN is proposed. In this approach, two methods are used for the premise of the accuracy and time of testing. The proposed model arranges the data records into two classes in SVM as well as ANN for further anaysis as shown in fig. The Back Propagation Neural Network (BPNN) with classification method is introduced, where the hypertension gene sequence is generated and then, thereafter the exact gene sequence. The performance of the BPNN techniques has been measured in the training phase as well as the testing phase with the various numbers of samples. The accuracy of this technique has improved in correspondence to the number of records. The feature selection plays a prominent role in the prediction of heart disease. ANN with back propagation is proposed for better prediction of the disease. The results obtained from the application of ANN are highly accurate and very precise.[3]

Chapter 4

Conclusion

Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible. Further extension of this study is highly desirable to direct the investigations to real world datasets instead of just theoretical approaches and simulations. The proposed hybrid HRFLM approach is used combining the characteristics of Random Forest (RF) and Linear Method (LM). HRFLM proved to be quite accurate in the prediction of heart disease. The future course of this research can be performed with diverse mixtures of machine learning techniques to better prediction techniques. Furthermore, new feature selection methods can be developed to get a broader perception of the significant features to increase the performance of heart disease prediction.[1]

Bibliography

[1] SPECIAL SECTION ON SMART CACHING, COMMUNICATIONS, COMPUTING AND CYBERSECURITY FOR INFORMATION-CENTRIC INTERNET OF THINGS.
Available at https://ieeexplore.ieee.org/document/8740989
Received May 13, 2019, accepted June 9, 2019, date of publication June 19, 2019, date of current version July 3, 2019.

[2] Article on Heart Disease Prediction Using Machine Learning.
Available at https://www.geeksforgeeks.org/disease-prediction-using-machine-learning/

[3] Article on Random Forest Regression using Python.
Available at https://www.geeksforgeeks.org/random-forest-regression-in-python/