Heart Disease prediction using Machine learning and Data Mining Technique

Conference Paper · March 2016

DOI: 10.090592/IJCSC.2016.018

3 authors, including:

Samir B Patel
Pandit Deendayal Energy University
90 PUBLICATIONS **1,468** CITATIONS

SEE PROFILE

IJCSC 0973-7391 Volume 7 • Number 1 Sept 2015 - March 2016 pp.129-137 available online at www.csjournalss.com

Heart Disease Prediction Using Machine learning and Data Mining Technique

Jaymin Patel, Prof.TejalUpadhyay, Dr. Samir Patel
Department of Computer Science and Engineering, Nirma University, Gujarat, India
Principal, Grow More Faculty of Engineering, Ahmedabad, Gujarat, India

Abstract—Heart disease is the main reason for death in the world over the last decade. Almost one person dies of Heart disease about every minute in the United States alone. Researchers have been using several data mining techniques to help health care professionals in the diagnosis of heart disease. However using data mining technique can reduce the number of test that are required. In order to reduce number of deaths from heart diseases there have to be a quick and efficient detection technique. Decision Tree is one of the effective data mining methods used. This research compares different algorithms of Decision Tree classification seeking better performance in heart disease diagnosis using WEKA. The algorithms which are tested is J48 algorithm, Logistic model tree algorithm and Random Forest algorithm. The existing datasets of heart disease patients from Cleveland database of UCI repository is used to test and justify the performance of decision tree algorithms. This datasets consists of 303 instances and 76 attributes. Subsequently, the classification algorithm that has optimal potential will be suggested for use in sizeable data. The goal of this study is to extract hidden patterns by applying data mining techniques, which are noteworthy to heart diseases and to predict the presence of heart disease in patients where this presence is valued from no presence to likely presence.
Keywords: Data Mining; Decision Support System; Health care; Health records; Classification.

I. INTRODUCTION

Heart disease is the leading cause of death in the world over the past 10 years (World Health Organization 2007). The European Public Health Alliance reported that heart attacks, strokes and other circulatory diseases account for 41% of all deaths (European Public Health Alliance 2010). Several different symptoms are associated with heart disease, which makes it difficult to diagnose it quicker and better. Working on heart disease patients databases can be compared to real-life application. Doctors knowledge to assign the weight to each attribute. More weight is assigned to the attribute having high impact on disease prediction. Therefore it appears reasonable to try utilizing the knowledge and experience of several specialists collected in databases towards assisting the Diagnosis process. It also provides healthcare professionals an extra source of knowledge for making decisions.

The healthcare industry collects large amounts of health-care data and that need to be mined to discover hidden information for effective decision making. Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amount of patients' data from which to extract useful knowledge, researchers have been using data mining techniques to help health care professionals in the diagnosis of heart disease (Helma, Gottmann et al. 2000). Data mining is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistical methods (Lee, Liao et al. 2000). Thus data mining refers to mining or extracting knowledge from large amounts of data. Data mining applications will be used for better health policy-making and prevention of hospital errors, early detection, prevention of diseases and preventable hospital deaths (Ruben 2009). Heart disease prediction system can assist medical professionals in predicting heart disease based on the clinical data of patients [1]. Hence by implementing a heart disease prediction system using Data Mining techniques and doing some sort of data mining on various heart disease attributes, it can able to predict more probabilistically that the patients will be diagnosed with heart disease. This paper presents a new model that enhances the Decision Tree accuracy in identifying heart disease patients. It uses the different algorithm of Decision Trees.

A. LITERATURE REVIEW

Prediction of heart disease using data mining techniques has been an ongoing effort for the past two decades. Most of the papers have implemented several data mining techniques for diagnosis of heart disease such as Decision Tree, Naive Bayes, neural network, kernel density, automatically defined groups, bagging algorithm and support vector machine showing different levels of accuracies (Yan, Zheng et al. 2003; Andreeva 2006; Das, Turkoglu et al. 2009; Sitar-Taut, Zdrenghea et al. 2009; Raj Kumar and Reena 2010; Srinivas Rani et al. 2010) on multiple databases of patients from around the world.

One of the bases on which the papers differ are the selection of parameters on which the methods have been used. Many authors have specified different parameters and databases for testing the accuracies. In particular, researchers have been investigating the application of the Decision Tree technique in the diagnosis of heart disease with considerable success. Sitair-Taut et al. used the Weka tool to investigate applying Naive Bayes and J48 Decision Trees for the detection of coronary heart disease. Tu et al. used the bagging algorithm in the Weka tool and compared it with J4.8 Decision Tree in the diagnosis of heart disease. In [9], the decision making process of heart disease is effectively diagnosed by Random forest algorithm. In [10] based on the probability of decision support, the heart disease is predicted. As a result the author concluded that decision tree performs well and sometimes the accuracy is similar in Bayesian classification.

In year 2013, S. Vijiyaraniet. al. [2] performed a work, An Efficient Classification Tree Technique for Heart Disease Prediction. This paper analyzes the classification tree techniques in data mining. The classification tree algorithms used and tested in this paper are Decision Stump, Random Forest and LMT Tree algorithm. The objective of this research was to compare the outcomes of the performance of different classification techniques for a heart disease dataset.

II. BACKGROUND

Millions of people are getting some sort of heart disease every year and heart disease is the biggest killer of both men and women in the United States and around the world. The World Health Organization (WHO) analysed that twelve million deaths occurs worldwide due to Heart diseases. In almost every 34 seconds the heart disease kills one person in world.

Medical diagnosis plays vital role and yet complicated task that needs to be executed efficiently and accurately. To reduce cost for achieving clinical tests an appropriate computer based information and decision support should be aided. Data mining is the use of software techniques for finding patterns and consistency in sets of data. Also, with the advent of data mining in the last two decades, there is a big opportunity to allow computers to directly construct and classify the different attributes or classes.
Learning of the risk components connected with heart disease helps medicinal services experts to recognize patients at high risk of having Heart disease. Statistical analysis has identified risk factors associated with heart disease to be age, blood pressure, total cholesterol, diabetes, hyper tension, family history of heart disease, obesity and lack of physical exercise, fasting blood sugar etc [3].

Researchers have been applying different data mining Techniques to help medicinal services experts with progressed exactness in the judgement of heart disease. Neural network, Naive Bayes, Decision Tree etc. are some techniques used in the diagnosis of heart disease.

Applying Decision Tree techniques has shown useful accuracy in the diagnosis of heart disease. But assisting health care professionals in the diagnosis of the world's biggest killer demands higher accuracy. Our research seeks to improve diagnosis accuracy to improve health outcomes.

Decision Tree is one of the data mining techniques that cannot handle continuous variables directly so the continuous attributes must be converted to discrete attributes. Couple of Decision Tree use binary discretization for continuous-valued features. Other important accuracy improving is applying reduced error pruning to Decision Tree in the diagnosis of heart disease patients. Intuitively, more complex models might be expected to produce more accurate results, but which techniques is best? Seeking to thoroughly investigate options for accuracy improvements in heart disease diagnosis this paper systematically investigates comparing multiple classifiers decision tree technique.

This research uses Waikato Environment for Knowledge Analysis (WEKA).The information of UCI repository regularly introduced in a database or spreadsheet. In order to use this data for WEKA tool, the data sets need to

be in the ARFF format (attribute-relation file format). WEKA tool is used for to pre-process the dataset. After reviewing all these 76 different attributes, the unimportant attributes is dropped and only the important attributes (i.e. 14 attributes in this case) is considered for analysis to yield more accurate and better results. The 14th one is basically a predicted attribute, which is referred as Class. With thorough comparison between different decision tree algorithms within WEKA tool and deriving the decisions out of it, would help the system to predict the likely presence of heart disease in the patient and will definitely help to diagnose heart disease well in advance and able to cure it in right time.

III. APPROACH AND METHODOLOGY

The following objectives are set for this heart prediction system.

- The prediction system should not assume any prior knowledge about the patient records it is comparing.
- The chosen system must be scalable to run against large database with thousands of data.

This chosen approach is implemented using WEKA tool. WEKA is an open source software tool which consists of an accumulation of machine learning algorithms for Data Mining undertakings. It contains apparatuses for information pre-processing, classification, regression, clustering, association rules, and visualization [4]. For testing, the classification tools and explorer mode of WEKA are used. Decision Tree classifiers with Cross Validation 10-fold in Test mode is considered for this study.
The following steps are performed in WEKA.

- Start the WEKA Explorer.
- Open CSV dataset file and save in ARFF format
- Click on Classify tab and select J48 etc (from Trees)
- from choose button.
- Select appropriate Test mode option.
- Click on Start button and result will be displayed

Data

For comparing various Decision Tree classification techniques, Cleveland dataset from UCI repository is used, which is available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease. The dataset has 76 attributes and 303 records. However, only 13 attributes are used for this study & testing as shown in Table 1.

Table 1: SELECTED HEART DISEASE ATTRIBUTES

Name	Type	Description
Age	Continuous	Age Age in years
Sex	Discrete	0 = female 1 = male
Cp	Discrete	Chest pain type: 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain 4 =asymptom
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar>120 mg/dl: 1-true 0=False
Exang Continuous Maximum heart rate achieved	Discrete	Exercise induced angina: 1 = Yes 0 = No
Thalach	Continuous	Maximum heart rate achieved
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3 = down sloping
Ca	Continuous	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
Thal	Discrete	3 = normal 6 = fixed defect 7= reversible defect
Class	Discrete	Diagnosis classes: 0 = No Presence 1=Least likely to have heart disease 2= >1 3= >2 4=More likely have heart disease

This paper has emphasized specifically on decision tree classifiers for heart beat prediction within WEKA. Decision tree was considered here among all types of Data mining techniques due to these below reasons. Decision tree filters are easy to implement and easy to understand. It is a method commonly used in data mining.Decision tree is one of thedata mining techniques demonstrating extensive achievementwhen contrasted with other data mining techniques. Itis a decision support system that uses a tree-like graphdecisions. Decision trees are the most powerful approachesin knowledge discovery and data mining. Decision trees arehighly effective tools in many areas such as data and textmining, information extraction, machine learning, and patternrecognition. It can handle input data like Nominal, Numeric &Text. It is able to process erroneous datasets or missing values.
A Decision Tree is used to learn a classification function which concludes the value of a dependent attribute (variable) given the values of the independent (input) attributes. This verifies a problem known as supervised classification because the dependent attribute and the counting of classes (values) are given [4]. Tree complexity has its effect on its accuracy. Usually the tree complexity can be measured by a metrics that contains: the total number of nodes, total number of leaves, depth of tree and number of attributes used in tree construction. Tree size should be relatively small that can be controlled by using a technique called pruning [5].

Univariate decision tree approach will be used here. In this technique, splitting is performed by using one attribute at internal nodes. This study can able to distinguish the dominant attributes and provides different labels of LIKELY PRESENCE for heart disease. In this paper, three decision tree algorithms namely J48 algorithm, logistic model tree algorithm and Random Forest decision tree algorithm are used for comparison. The proposed methodology involves reduced error pruning, confident factor and seed parameters to be considered in the diagnosis of heart disease patients. Reduced error pruning has shown to drastically improve decision tree performance. These three decision tree algorithms are then tested to identify which combination will provide the best performance in diagnosing heart disease patients.
A correlation is based on affectability, specificity and precision by genuine positive and false positive in confusion matrix. To have a reasonable correlation between these algorithms, preparing time in seconds and tree size proportion for every system is considered with 10-fold stratified cross validation. The general approach took after for Decision Tree classification for satisfying the objective is:

Training => Algorithm => Model => Testing => Evaluation

Classification Tree Algorithms Used
J48 algorithm:

J48 is an open source Java implementation of the C4.5 algorithm in the WEKA tool. This algorithm utilizes an avaricious method to make decision trees for classification and uses decreased-error pruning [6]. Decision tree is built by examining information hubs, which are utilized to assess hugeness of existing highlights.J48 algorithm is an extension of ID3 algorithm and possibly creates a small tree. It uses divide and conquers approach to growing decision trees [7].At every node of the tree, the algorithm picks a attribute that can further part the samples into subsets. Every leaf node speaks to a class or decision.

Basic steps to construct tree are

- Check whether all cases belongs to same class, then the tree is a leaf and is labeled with that class.
- For each attribute, calculate the information and information gain.
- Find the best splitting attribute (depending upon current selection criterion).

J48 with Reduced error Pruning:

Pruning is very important technique to be used in tree creation because of outliers. It also addresses overfitting. Datasets may contain little subsets of instances that are not well defined. To classify them correctly, pruning can be used. Separate and Conquer rule learning algorithm is basis to prune any tree. This rule learning scheme starts with an empty set of rules and the full set of training instances. Reduced-error pruning is one of such separate and conquer rule learning scheme. There are two types of pruning i.e.

- Post pruning (performed after creation of tree)
- Online pruning (performed during creation of tree).

After extracting the decision tree rules, reduced error pruning was used to prune the extracted decision rules. Reduced error pruning is one of the fastest pruning methods and known to produce both accurate and small

decision rules (Esposito, Malerba et al. 1997). Applying reduced error Pruning provides more compact decision rules and reduces the number of extracted rules.

The run-time complexity of J48 algorithm matches to the tree depth which is linked to tree size and number of examples. So their greatest disadvantage is size of J48 trees, which increases linearly with the number of examples. J48 rules slow for large and noisy datasets. Space complexity is very large as we have to store the values repeatedly in arrays.

Logistic Model Tree Algorithm:

Logistic Model Tree is the classifier for building logistic model trees, which consist of a decision tree structure with logistic regression function at the leaves. The algorithm can oversee parallel and multi-class target variables, numeric and nominal attributes and missing qualities [8]. A combination of learners that rely on simple regression models if only little and/or noisy data is available and add a more complex tree structure if there is enough data to warrant such a structure. LMT uses cost-complexity pruning. This algorithm is significantly slower than the other algorithms.

As in decision tree, the tested attributes is associated with every inner node. The attributes with k values, the node has k child nodes for nominal attributes and depending on the value of the attribute, the instances are sorted down. For the attributes of numeric, the node has two child nodes and comparing the attributes of tested value to a threshold (the instances are sorted down based on threshold [9].

Logistic Model Trees have been demonstrated to be extremely exact furthermore, smaller classifiers in diverse examination regions. Their most noteworthy weakness is the computational unpredictability of inciting the logistic regression models in the tree. Anyway the prediction of a model is acquired by sorting it down to a leaf what's more, utilizing the logistic prediction model connected with that leaf. A solitary logistic model is less demanding to translate than J48 trees. However fabricating LMT'stake longer time. It can likewise be demonstrated that trees produced by LMT are much littler than those produced by J48.

To construct a logistic model tree by developing a standard classification tree, building logistic regression models for all node, pruning a percentage of the sub-trees utilizing a pruning model, and combining the logistic models along a way into a solitary model in some manner is performed.

The pruning plan uses cross-validation to get more steady pruning results. In spite of the fact that this expanded the computational multifaceted nature, it brought about littler and for the most part more accurate trees. These thoughts lead to the following algorithm for developing logistic model trees:

Tree developing begins by building a logistic model at the root utilizing the LogitBoost algorithm. The quantity of cycles (and basic relapse capacities fmj to add to Fj) is resolved utilizing 10 fold cross-validation. In this process the information is part into preparing and test set 10 times, for each preparation set LogitBoost is rush to a greatest number of cycles and the lapse rates on the test set are logged for each cycle and summed up over the distinctive folds. The quantity of emphases that has the least whole of blunders is utilized to prepare the LogiBoost algorithm on all the information.
This gives the logistic regression model at the base of the tree.

Like other tree impelling systems, LMT does not oblige any tuning of parameters. LMT produces a solitary tree containing double parts on numeric properties, multi-route parts on ostensible ones and logistic regression models atthe leaves, and the algorithm guarantees that just applicable attributes are incorporated in the last.

Random Forest Algorithm:

Random forest is an ensemble classifier that consists of many decision trees. The output of the classes is represented by individual trees. It is derived from random decision a forest that was proposed by Tin Kam Ho of Bell Labs in 1995 [9]. This method combines with random selection of features to construct a decision trees with controlled variations. The tree is constructed using algorithm as discussed.

Let N be the number of training classes and M be the number of variables in classifier.

- The input variable m is used to determine the node of the tree. Note that m<M.
- Choosing n times of training sets with the replacement of all available training cases N by predicting the classes, estimate the error of the tree.
- Choose m variable randomly for each node of the tree and calculate the best split.

- At last the tree is fully grown and it is not pruned. The tree is pushed down for predicting a new sample. When the terminal node ends up, the label is assigned the training sample. This procedure is iterated over all trees and it is reported as random forest prediction.

Multi-classifiers are the aftereffect of joining a few individual classifiers. Troupes of classifiers towards expanding the execution have been presented. [5].

Random Forest (RF) is one of the case of such procedures. RF as a multi classifier formed by choice trees where each tree ht had been created from the set of information preparing and a vector θ t of arbitrary numbers indistinguishably disseminated and free from the vectors. Vectors θ1, θ 2 ,.., θ t-1 used to create the classifiers h1; h2; :::; ht-1 . Every decision tree is manufactured from random subset of the preparation dataset. It utilized a random vector that is produced from some altered likelihood dissemination, where the likelihood circulation is shifted to centre samples that are difficult to arrange. A Random vector can be joined into the tree-becoming process from various perspectives. The leaf hubs of each one tree are named by evaluations of the back dissemination over the information class names. Every interior hub contains a test that best parts the space of data to be arranged. Another, concealed occasion is ordered by sending it down every tree and conglomerating the arrived at
leaf appropriations.

There are three methodologies for Random Forest, for example, Forest-RI(Random Input choice) and Forest-RC (Random blend) and blended of Forest-RI and Forest-RC.

The Random Forest procedure has some desirable qualities, for example

- It is not difficult to utilize, basic and effortlessly parallelized.
- It doesnt oblige models or parameters to choose aside from the quantity of indicators to pick at arbitrary at every node.
- It runs effectively on extensive databases; it is moderately strong to anomalies and commotion.
- It can deal with a huge number of information variables without variable deletion; it gives evaluations of what variables are important in classification.
- It has a successful system for assessing missing information and keeps up accuracy when a vast extent of the data are missing, it has methods for adjusting error in class populace unequal data sets.

Evaluation of Classification Algorithms

The execution of Classification algorithm is generally analysed by assessing the affectability, specificity, and accuracy of the classification. The sensitivity is proportion of positive instances that are correctly classified as positive (i.e. the proportion of patients known to have the disease, who test positive for it).The specificity is the proportion of negative instances that are correctly classified as negative (i.e. theproportion of patients known not to have the disease, who testnegative for it).The accuracy is the proportion of instances that are correctly classified. To quantify the dependability of the execution of proposed model, the information is isolated into preparing and testing data with 10-fold stratified cross validation these values are defined as,

Sensitivity = True Positive/(True Positive + False Negative)
Specificity = True Negative/(True Negative +False Positive)
Accuracy = (True Positive + True Negative) / (True
Positive + True Negative+ False Negative+ False Positive)

All measures can be ascertained focused around four qualities specifically True Positive, False Positive, False Negative, and False Positive where,

- True Positive (TP) is various effectively classified thatan instances positive.
- False Positive (FP) is a number of incorrectly classifiedthat an instance is positive.
- False Negative (FN) is a number of incorrectly classifiedthat an instance is negative.
- True Negative (TN) is a number of correctly classifiedthat an instance is negative.
- F-Measure is a way of combining recall and precisionscores into a single measure of performance.
- Recall is the ratio of relevant instances found in thesearch result to the total of all relevant instances.
- Precision is the proportion of relevant instances in the results returned.
- Receiver Operating Characteristics (ROC) Area is atraditional to plot this same information in a normalizedform with 1-false negative rate plotted against thefalse positive rate.
- For every algorithm, the test choice cross-validationwere utilized. As opposed to holding a part for testing,the cross-validation repeats the training and testingprocess a few times with random forest

samples. Thestandard for this is 10-fold cross-validation. The datais partitioned arbitrarily into 10 sections in which theclasses are represented in the same proportions asin the full dataset(stratification).Each one section is held out thus and the algorithm is trained on the nineremaining parts; then its error rate is computed onthe holdout set. At long last, the 10 error estimatesare found the middle value of to yield an overallerror estimate. For J48 and Random Forest, all the testswere run with ten different random seeds. Choosing thedifferent random seeds is carried out to normal outstatistical variations.

IV. RESULTS

The decision tree classification was performed using J48algorithm, logistic model trees algorithm and Random Forestalgorithm on UCI repository. The experimental results isunder the framework of WEKA 3.6.10. All experiment wereperformed on Core I3 with 2.4GHz CPU and 4GB RAM. Theexploratory results are divided into a few sub thing for lessdemanding examination and assessment.

A. J48 with ReducedErrorpruning Algorithm

The sample of J48 algorithm is connected on UCIrepository and the confusion matrix is produced for classhaving 5 conceivable qualities are demonstrated in Fig 2. Theconfusion matrix is imperative viewpoint to be considered.From this matrix, classifications can be made. Theresults ofthe J48 algorithm are indicated in Table 2.

————Confusion Matrix————

a	b	c	d	e	
146	8	4	6	0	|a = 0
31	9	9	6	0	|b = 1
9	5	13	8	1	|c = 2
11	7	10	4	3	|d = 3
2	5	3	3	0	|e = 4

TABLE II. CLASSIFICATION RESULT FOR J48

	Train Error	Test Error
J48	0.1423221	0.1666667

J48 model sacrifices error rate for a clearer decision processand as a result the error is acceptable.

B. Logistic Model Tree Algorithm

The example of logistic model trees algorithm is connectedon UCI repository and the confusion matrix is produced forclass gender having two conceivable qualities are indicatedin Fig 4. The results of LMT algorithm are demonstrated inTable 3.

————Confusion Matrix————

a	b	c	d	e	
148	12	2	1	1	|a = 0
31	10	6	8	0	|b = 1
8	12	4	10	2	|c = 2
4	11	11	7	2	|d = 3
0	5	2	6	0	| e = 4

TABLE IIICLASSIFICATION RESULT FOR LOGISTIC MODEL TREE ALGORITHM

	Train Error	Test Error
Logistic Model Tree Algorithm	0.1156716	0.137931

C. Random Forest algorithm

The example of Random Forest algorithm is connected onUCI repository and the confusion matrix is created for classhaving 5 qualities are demonstrated in Fig 5. The result ofthe Random Forest algorithm are demonstrated in Table 4.

————Confusion Matrix————

a	b	c	d	e	
152	7	2	3	0	|a = 0
34	4	10	5	2	|b = 1
10	11	7	7	1	|c = 2
5	11	12	5	2	|d = 3
1	5	2	3	2	| e = 4

TABLE IV. CLASSIFICATION RESULT FOR RANDOM FOREST ALGORITHM

	Train Error	Test Error
Random Forest Algorithm	0	0.2

As is mentioned above, we use random forest to choose keyvariables to project our data on. Since the model is flexible, the0 train error is explainable while the 0.2 test error is acceptable.

As is mentioned above, we use random forest to choose keyvariables to project our data on. Since the model is flexible, the0 train error is explainable while the 0.2 test error is acceptable.

V. COMPARISON OF METHODOLOGIES

TABLE V. COMPARISON OF DIFFERENT ALGORITHM RESULTS

	J48	Logistic Model Tree Algorithm	Random Forest Algorithm
Train Error	0.1423221	0.1656716	0
Test Error	0.1666667	0.237931	0.2

When comparing the results with LMT and Random Forestalgorithm, J48 algorithm achieved higher sensitivity andaccuracy while LMT achieved higher specificity than J48 andRandom Forest algorithm. So overall from Table 10 and Table11, it is concluded that J48 (with ReducedErrorPruning) hasgot the best overall performance.

Also, J48 algorithm utilization reduced-error pruning form lessnumber of trees. The LMT algorithm manufactures the littlesttrees. This could show that cost-many-sided quality pruningprunes down to littler trees than decreased lapse pruning, yetit additionally demonstrate that the LMT algorithm does nothave to assemble huge trees to group the information. TheLMT algorithm appears to perform better on data sets withnumerous numerical attributes, while for good execution for3 algorithm, the data sets with couple of numerical qualitiesgave a superior execution. We can see from the outcomes thatJ48 is the best classification tree algorithm among the threewith pruning system.

VI. CONCLUSION

By analysing the experimental results, it is concluded thatJ48 tree technique turned out to be best classifier for heartdisease prediction because it contains more accuracy and leasttotal time to build. We can clearly see that highest accuracybelongs to J48 algorithm with reduced error pruning followedby LMT and Random Forest algorithm respectively. Alsoobserved that applying reduced error pruning to J48 results inhigher performance while without pruning, it results in lower Performance. The best algorithm J48 based on UCI data hasthe highest accuracy i.e. 56.76% and the total time to buildmodel is 0.04 seconds while LMT algorithm has the lowestaccuracy i.e 55.77% and the total time to build model is 0.39seconds.

In conclusion, as identified through the literature review, webelieve only a marginal success is achieved in the creation ofpredictive model for heart disease patients and hence thereis a need for combinational and more complex models toincrease the accuracy of predicting the early onset of heartdisease.

VII. FUTURE WORK

There are many possible improvements that could beexplored to improve the scalability and accuracy of thisprediction system. Due to time limitation, the following research/
work needs to be performed in the future.

- Like to make use of testing different discretizationtechniques, multiple classifiers Voting technique anddifferent Decision tree types like information gain,gain ratio and Gini index. Eg. Experiment need to performon use of Equal Frequency Discretization GainRatio Decision Trees by applying nine Voting schemein order to enhance the accuracy and performance ofdiagnosis of heart disease.
- This paper proposes a framework using combinationsof support vector machines, logistic regression and decision trees to arrive at an accurate prediction ofheart disease. Further work involves development ofsystem using the mentioned methodology to be usefor checking the imbalance with other data miningmodels.
- Like to explore different rules such as Association, Clustering, K-means etc for better efficiency and easeof simplicity.
- To make use of Multivariate Decision Tree approachon smaller and larger amount of data.

REFERENCES:

[1]C. S. Dangare and S. S. Apte, "Improved study of heart disease predictionsystem using data mining classification techniques," InternationalJournal of Computer Applications, vol. 47, no. 10, pp. 44–48, 2012.
[2] S. Palaniappan and R. Awang, "Intelligent heart disease prediction systemusing data mining techniques," pp. 108–115, 2008.
[3] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, "Hybrid intelligent modelling schemes for heart disease classification," Applied Soft Computing,vol. 14, pp. 47–52, 2014.
[4] M. Shouman, T. Turner, and R. Stocker, "Using data mining techniquesin heart disease diagnosis and treatment," pp. 173–177, 2012.;3
[5] P. V. Ankur Makwana, "Identify the patients at high risk of re-admissionin hospital in the next year," International Journal of Science andResearch, vol. 4, pp. 2431–2434, 2015.
[6] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, "Computationalintelligence for heart disease diagnosis: A medical Knowledge driven approach," Expert Systems with Applications, vol. 40, no. 1, pp. 96–104,2013.
[7] Y. Xing, J. Wang, Z. Zhao, and Y. Gao, "Combination data miningmethods with new medical data to predicting outcome of coronary heartdisease," pp. 868–872, 2007.
[8] Combination data mining methods with new medical data to predictingoutcome of coronary heart disease," in Convergence InformationTechnology, 2007. International Conference on. IEEE, 2007, pp. 868–872.
[9] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, "Hybrid intelligent modelling, schemes for heart disease classification," Applied Soft Computing,vol. 14, pp. 47–52, 2014.