7th National Conference on
"Role of Engineers in Nation Building" organized by VIVA Institute of Technology, Mumbai
(1st and 2nd March 2019)
International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

Survey on An Innovative Approach for Cardiovascular Heart Disease Prediction Using Machine Learning

Sankita Hate[1], Tejal Patil[2], Praneeta Pitale[3], Tatwadarshi Nagarhalli[4]
Department of Computer Engineering, Mumbai University, India

Abstract— *Artificial Intelligence (AI) is an area of computer science that emphasizes the creation of intelligent machine that work and react like human. Predicting the chances of heart attack according to person's symptoms using Gradient boosting and C4.5 algorithm, Neural network, support vector machine, Random forest, Naïve Bayes. Disease diagnosis is the process to find the disease with specified details of a person's symptoms. The healthcare industry collects a huge amount of data which is not properly mined and not put to the optimum use. Our research focuses on this aspect of Medical diagnosis by learning pattern through the collected data of diabetes, hepatitis and heart diseases and to develop intelligent medical decision support systems to help the physicians. The purpose of this paper is to develop a cost effective treatment using data mining technologies for facilitating data base decision support system. In this paper using varied data mining technologies an attempt is made to assist in the diagnosis of the disease in question. Decision tree is one of many data mining analytical tools that can be utilized to make predictions for medical data. From the study it is observed that Gradient boosting Algorithm improves the accuracy of the heart disease prediction system.*

Keywords— Artificial Intelligence, Data Mining, Gradient boosting and C4.5 algorithm.

I. INTRODUCTION

Machine learning is building model to quickly analyze data and deliver results both historical and real –time data. With machine learning, healthcare service provider can make better decisions on patient's diagnosis and treatment options, which leads to the overall improvement of healthcare services. The Healthcare industry generally clinical diagnosis is done mostly by doctor's expertise and experience. Computer Aided Decision Support System plays a major role in medical field. With the growing research on heart disease predicting system, it has become important to categories the research outcomes and provides readers with an overview of the existing heart disease prediction techniques in each category.

Diagnosis is intricate process that should be precisely and intensively executed. Decision tree is one of many data mining analytical tools that can be utilized to make predictions for medical data. From the study it is observed that Gradient boosting Algorithm improves the accuracy of the heart disease prediction system. Disease diagnosis is the process to find the disease with specified details of a person's symptoms. Diagnosing the Disease is time consuming due to the need to analysis relevant microorganisms. Due to large growth in world's population, Classification model receives a great deal in any domain of research and also a consistent tool for medical disease diagnosis.

II. LITERATURE SURVEY

2.1 A Fast Correlation Filter Based Gradient Boosting Classifier For Disease Diagnosis [1].

The main objective of the FCF-GBC technique is effectively performs disease diagnosis with two processing steps. FCF uses symmetrical uncertainty to calculate the dependences of attributes and discovers the relevant attributes. After that, A Gradient Boosting Classifier is used for classifying and predicting the heart and stroke disease from the extracted attributes. This helps to improve the classification accuracy with minimum time. Gradient boosting is a machine learning technique used for classification as well as prediction approach. The structure of the gradient boosting is the ensemble of weak prediction, generally using as decision trees. The GBC is iteratively learning an ensemble of weak classifier and combining them into a final strong classifier to provide the final results for disease diagnosing.

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

2.2 Heart disease prediction system based on hidden naive Bayes classifier [2].

To overcome the drawbacks of Naïve Bayes, Hidden Naïve Bayes classifier is proposed. In this paper they have used the heart catalogue dataset in ARFF from the VCP repository, then adopted the following pre-processing technique to run the experiment. Replace missing search (mean) , Discretization (Numeric attributes were discretized), Enter quartile range filters (IQR) They used accuracy, sensitivity, specificity, positive predictive value measures to evaluate the performance of HNB. Then confusion matrix is used to visualize performance of an algorithm. It is used to measure the incorrect and correct prediction made by the classifier.

2.3 Heart Attack prediction system [4].

The implementation of this prototype uses the abstract obtained from VCL'S machine learning repository rapid miner used to cleaning the dataset. Rapid miner used to establish the best fitting algorithm for the given dataset and an algorithm including Naïve Bayes Decision trees, K-Nearest and Random Forest were compared by building their process in rapid mine. The outcome of this activity showed that Naïve Bayes gives the highest accuracy on the given dataset. The dataset of 14 attributes in total out of which 13 are predictor various and one feature is a binary response variable. It started with uploading the dataset into the design section of rapid miner using the "retrieve" operator followed by the use of "set ratio" operator for specifying the class label and finally the "replace missing value" operator to handle the missing values by substituting with the mean. An interactive web interface was developed to access the classifier the choke the risk factor. Once the model makes a prediction, it is displayed to the user.

2.4 Prediction of Heart Disease Using Neural Network [5].

In this paper, a heart disease prediction system which uses artificial neural network back propagation algorithm is proposed. The neural network was trained with back propagation algorithm to predict absence or presence of heart disease with accuracy of 95%. The dataset was split into three parts: training, testing and validation. Then training was done with Back propagation Algorithm. After training process, the performance of the proposed system was computed by testing the neural network with test data by different metrics including accuracy, precision and recall.

2.5 Prediction of heart disease using hybrid technique for selecting features [6].

The proposed method is based on the selecting features that involves two stage feature selection by combination gain ratio and recursive feature elimination SVM algorithm. In this paper feature selection along with random forest and naïve Bayes classifier as well as it consist of two classes either positive or negative result. The main motive of this research is to classify the data for effective decision making in field of medical, advance technique of data mining technique are used. Feature selection approach is employed to remove irrelevant or redundant features. More specifically gain ratio and SVM recursive feature elimination algorithm are applied to dataset. Combine results are used to evaluate subset of features.

2.6 Analytical study of heart disease prediction comparing with different algorithm [7].

In this paper machine learning data mining technique are used for diagnosis heart disease. Genetic algorithm, particle swarm optimizations and artificial neural network are used for predicting heart disease to extracting the knowledge information from large dataset. Input data are trained and several classification techniques are applied so that they extract hidden useful information. It has capability to handle difficult problem. It is more robust and has property of inherent parallelism. Genetic algorithm follows "survival fitness" principal. It is helpful for solving complex design optimization problem and solve discrete, continuous variable without gradient information.

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

2.7 Heart Function Monitoring, Prediction and Prevention of Heart Attacks: Using Artificial Neural Networks [11].

In this paper we have developed an efficient method to acquire the clinical and ECG data, so as to train the Artificial Neural Network to accurately diagnose the heart and predict abnormalities if any. The overall process can be categorized into three steps. Firstly, we acquire the ECG of the patient by standard 3 lead pre jelled electrodes. The acquired ECG is then processed, amplified and filtered to remove any noise captured during the acquisition stage. This analog data is now converted into digital - format by A/D converter, mainly because of its uncertainty. Secondly we acquire 4-5 relevant clinical data's like mean arterial pressure (MAP), fasting blood sugar (FBS), heart rate (HR), cholesterol (CH), and age/gender. Finally we use these two data's i.e. ECG and clinical data to train the neural network for classifying the heart disease and to predict abnormalities in the heart or it's functioning.

2.8 Analytical Study of Heart Disease Diagnosis Using Classification Techniques [12].

This disease mostly affected in male because smoking habits. This paper analyses the different kinds of heart disease using the classification techniques. In this paper the potential of nine (9) classification techniques was evaluated of prediction of heart disease. Namely decision tree, naive Bayesian neural network, SVM.ANN, KNN. My proposed algorithm of Apriority algorithm and SVM (support vectormachine) in heart disease prediction. Different types of data mining techniques available are classification, cluster, feature selection, association rule can be analyzing the heart disease prediction. Data mining basically using the 4 Techniques, Classification, Cluster, Feature selection D, Association rule

2.9 Prediction in heart disease using technique of data mining [13].

Data mining technique can help as remedy in this circumstances. The three main methods which are analysed on medical data set are Naive Bayes , neural network and decision tree .the target of this paper is to find out aspects of the use of healthcare data for aid of people by method of machine learning by further more data mining procedure. The main aim is to suggest an automated system for diagnosis of heart disease by taking into account earlier information and data.Data mining examination, machine learning engineering to extract hidden connections from substantial data characterizes data mining as "a nontrivial extraction of implied, la and possibly valuable data from that is stored in a database" characterizes data mining as "a choice, investigation and demons amounts of information to find relations with the point of getting c results for the manager of database".

III. RELATED WORKS

As a preliminary stage, the data set is pre-processed by using Numeric to Nominal and Replace Missing Value techniques. After cleaning, the data set is trained for accuracy the next stage is the extraction of redundant feature for prediction. This is to be effected by using the Swarm Intelligence Techniques hybrid with Rough set Algorithm. The data set is validated to acquire the optimal redundant feature for prediction. In practice, given some specific loss function (y, f) and/or a custom base-learner h(x, θ), the solution to the parameter estimates can be difficult to obtain. To deal with this, it was proposed to choose a new function h(x, θt) to be the most parallel to the negative gradient {gt(xi)}Ni = 1along the observed data.

$$gf(x) = Ey[[(\partial\Psi(y,f(x)) /\partial f(x))] |x]f(x)=f \text{ `f-1(x)}$$

Instead of looking for the general solution for the boost increment in the function space, one can simply choose the new function increment to be the most correlated with −gt(x). This permits the replacement of a potentially very hard optimization task with the classic least-squares minimization one [14].C4.5 is a software extension of the basic ID3 algorithm designed by Quinlan to address the following issues not dealt with by ID3: Avoiding over fitting the data, Determining how deeply to grow a decision tree, Reduced error pruning, Rule post-pruning, Handling continuous attributes. Example temperature, choosing an appropriate attribute selection measure, handling training data with missing attribute values, handling attributes with differing costs, improving computational efficiency [15].This algorithm also maximizes the correlation between the error of the whole network and the newly created neuron, which makes the comparison more evident.

7th National Conference on
"Role of Engineers in Nation Building" organized by VIVA Institute of Technology, Mumbai
(1st and 2nd March 2019)
International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

IV. PROPOSED SYSTEM

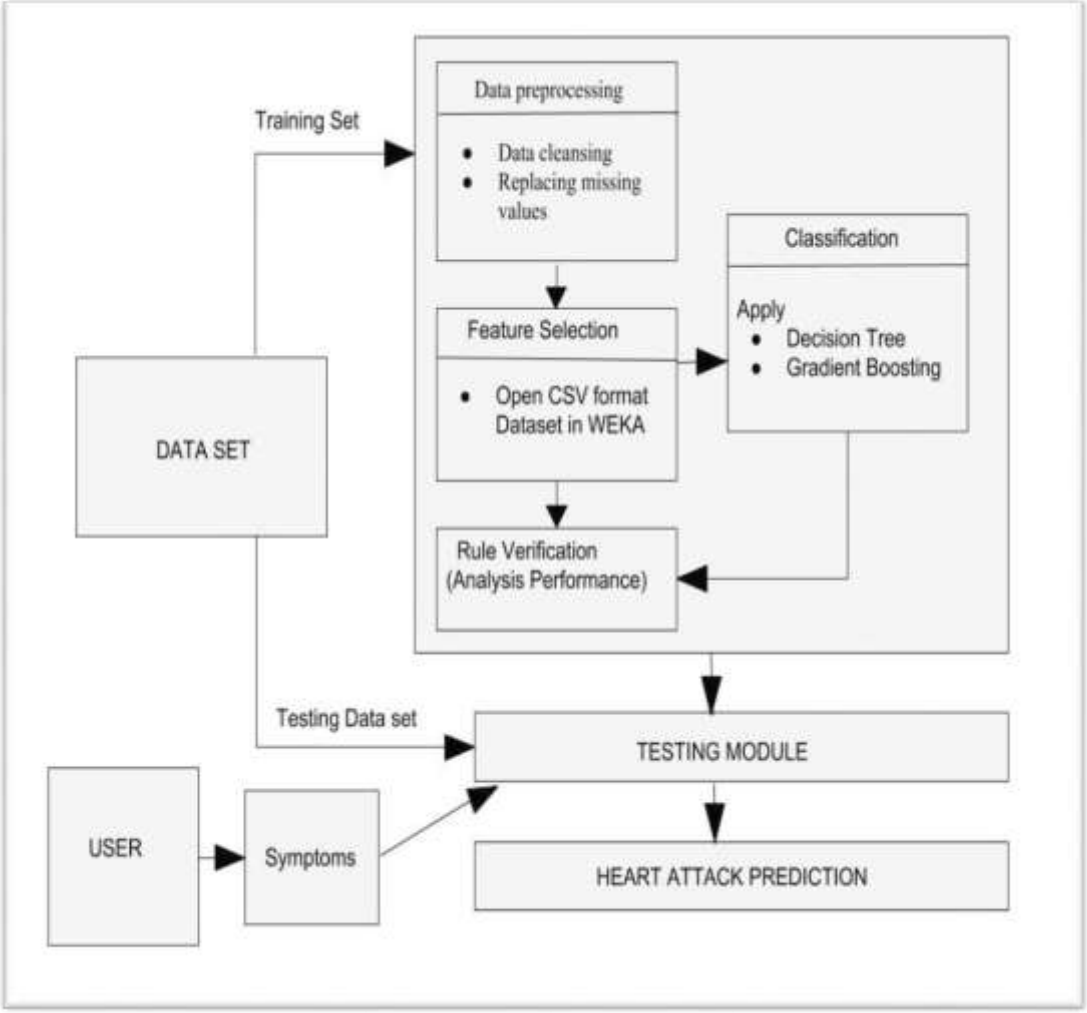

Fig 4.1: System flow

In Fig 4.1, Data set splits into two parts, one (training data set) goes into training module and another one (training set) goes to data processing, in data processing data is cleansed by replacing missing values. After data processing, Feature Selection is performed on data. Applying Gradient boost and decision tree algorithm in classification part. Then the processed data is passed to the testing module. According to the user's symptoms, system will perform prediction and result will be displayed to the user.

V. ANALYSIS TABLE

The below table is the summary of the studied research papers and the different techniques used on for prediction analysis of cardiovascular disease.

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

Table 1

Sr. No.	Title of Paper	Technique/ Methods	Disadvantages
1.	A Fast Correlation Filter Based Gradient Boosting Classifier For Disease Diagnosis [1].	Fast correlation Filter, Gradient Boosting Algorithm	It require carefull tuning and sequential nature.
2.	Heart disease prediction system based on hidden naive Bayes classifier [2].	Hidden naïve Bayes, data mining,	Needs more training time.
3.	Cardiovascular disease detection using new ensemble classifier [3].	regression; machine learning	It will not produce accurate result for large datasets.
4.	Heart Attack prediction system [4].	Back propagation algorithm, Multilayer Perceptron Neural Network	Large complexity of the networks structure
5.	Prediction of Heart Disease Using Neural Network [5].	Neural network ,Rough set, Naïve Bayes	These techniques gives the low accuracy.
6.	Prediction of heart disease using hybrid technique for selecting features [6].	Hybrid selecting feature	Hard to determine threshold value.
7.	Analytical study of heart disease prediction comparing with different algorithm [7].	Average K-nearest algorithm	It needs lots of data.
8.	Heart attack prediction system using average k-nearest neighbour algorithm [8].	Data mining (PSO, ANN)	Still too slow compared to classical approaches.
9.	Identification of heart failure by using unstructured data of cardiac patients [9].	Data mining	This will not give better performance. Accuracy was too low.
10.	Heart Function Monitoring, Prediction and Prevention of Heart Attacks: Using Artificial Neural Networks [10].	Artificial Neural Network	It consist of huge complexity.
11.	Analytical Study of Heart Disease Diagnosis Using Classification Techniques [11].	decision tree, naive Bayesian neural network, SVM.ANN, KNN. My proposed algorithm of Apriority algorithm and SVM (support vector machine)	Memory limitation.
12.	Prediction in heart disease using technique of data mining [12].	Naive Bayes, neural network and decision tree	There is an absence of successful analysis methods to find connections and patterns in health care data , low accuracy.

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

VI. CONCLUSION

The use of gradient boosting and C4.5 algorithm for heart disease prediction will improve the system performance and reduce the error rating. This model use global dataset from UCI repository of machine learning. The accuracy of the system can be improve with the implementation of other powerful ensemble method by using local data set. So, in the proposed system, machine learning algorithm are used for prediction analysis. Combine gradient boosted and C4.5 are two algorithm used, which gives better accuracy and performance as compared to previous approaches in Prediction Analysis Compared to other machine learning algorithm such as RF, SVM. it gives better results in terms of accuracy and performance.

REFERENCES

[1] K. S. Thirunavukkarsu, "A fast correlation filter based gradient boosting classifier for disease diagnosis", International Journal of Advance Research in Computer Science (IJARCS), 2017.

[2] S. Manikandan, "Heart attack prediction system", International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017.

[3] M. A. Jabber and S. Samreen, "Heart disease prediction system based on hidden naïve bayes classifier", International Conference on Circuits, Controls, Communications and Computing (I4C), 2016.

[4] H. A. Esfahani and M. G. hazanfari, "Cardiovascular disease detection using new ensemble classifier", IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), 2017.

[5] T. Karayılan and O. Kılıç, "Prediction of heart disease using neural network", International Conference on Computer Science and Engineering (UBMK), 2017.

[6] K. Pahwa and R. Kumar, "Prediction of heart disease using hybrid technique for selecting features", 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), 2017.

[7] C. Kalaiselvi, "Diagnosing of heart diseases using average k-nearest neighbour algorithm of data mining", 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016.

[8] S. Bharati and S. N. Singh, "Analytical study of heart disease Prediction comparing with different algorithm", International Conference on computing, communication and automation (ICCCA), 2015.

[9] M. Saqlain, W. Hussain, N. A. Saqib and M. A. Khan, "Identification of Heart Failure by Using Unstructured Data of Cardiac Patients", 45th International Conference on Parallel Processing Workshops (ICPPW), 2016.

[10] M. Panahiazai, V. Taslimitehrani, N. Pereira and J. Pathak, "Heart Function Monitoring Prediction and Prevention Of Heart Attack: Using Artificial Neural Network", International Conference On Contemporary Computing and Informatics (IC3), 2015.

[11] C. Sowmiya and P. Sumitr, "Analytical study of heart disease diagnosis using classification techniques", IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2017.

[12] M. Gandhi, "Prediction in heart disease using technique of data mining", International Conference on Futuristic Trends on Computational Analysis and Knowledge Management(ABLAZE), 2015.

[13] R. G. Saboji, "A scalable solution for heart disease prediction using classification mining technique", International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017