IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol-6, Special Issue-4, May 2018 E-ISSN: 2347-2693

A conceptual method to enhance the prediction of heart diseases using big data Techniques

R. Sharmila [1*], S. Chellammal [2]

[1*]Dept. of Computer Science, Bharathidasan University Constituent Arts & Science College, Tiruchirappalli, India
[2] Dept. of Computer Science, Bharathidasan University Constituent Arts & Science College, Tiruchirappalli, India

[*]*Corresponding Author:* sharmiparam@gmail.com,

Available online at: www.ijcseonline.org

***Abstract*—** As death rate due to heart diseases is increasing significantly, prediction of heart disease with required accuracy becomes a critical issue in health care industry. Data mining and machine learning algorithms, more specifically classification algorithm plays an important role in prediction. Still the accuracy of prediction is influenced by the evolving size of data, nature or format of data and velocity of data. Keeping these factors in mind, Big data based model has been proposed after a careful investigation on existing analytical algorithms. In this paper, some of the existing literature related to the prediction of heart diseases using data mining is presented. Inferences are drawn to find out the essential attributes to be considered for prediction. A study is carried out to find which algorithm will work better for prediction of heart diseases. With inference drawn, an approach is proposed based on Hadoop and MapReduce programming paradigm. It is proposed to employ Support Vector Machine(SVM) in parallel fashion in order to improve the accuracy of prediction. The overview of proposed model is presented.

***Keywords*—** Bid data in prediction, classification techniques in heart disease prediction, parallel SVM

I. INTRODUCTION

In medical science heart disease is one of the major challenges[1]. Several attributes which lead to heart disease include family history, smoking, cholesterol, high blood pressure, obesity and lack of physical exercise, etc [2]. Hence, there is a need to predict the heart disease in its earlier stage. But prediction of heart disease is difficult task in medical field [3]. Traditional way of prediction is mainly based on medical tests such as Elector Cardio Gram(ECG), Stress test, MRI, etc[4]**.** In addition to medical test, prediction using data mining techniques such as Decision Tree, Naive Bayes, Neural Network and Support Vector Machine yields better performance. However prediction of heart disease is an open issue. Several research works are being carried in this track. From literature survey it is understood that existing technologies and techniques have some limitations for example Evolving size of heart disease data demand alternate technologies such as Big data for better prediction. In recent days, the data for prediction exist in various varieties such as images, text, audio, picture, sms, mms,etc. But these data are rarely used to support clinical decision making [5].So, prediction requires an alternate technology which can handle any type of structured and unstructured data. In addition to size and format, the data can arise with speed. As traditional system are insufficient in handling data with speed, prediction of heart disease require yet other systems/databases which can handle the velocity.

In this context, the present work highlights the survey on existing research works on the prediction of heart disease. Further, it presents an alternate approach for the prediction of heart disease.

Rest of the paper is organized as follows. Section II highlights the survey related to the theme of the paper. Section III describes the inferences drawn from the survey. Section IV presents the proposed model for disease prediction using big data technologies and section V concludes research work with future directions.

II. LITERATURE SURVEY

Some survey has been carried in to find applicability of classification techniques such as Decision Tree, Naive Bayes algorithm, Neural Network and Support Vector Machine(SVM) for Heart disease prediction. The Details of survey are given in Table 1.

Table 1 Role of Classification Techniques in Heart Disease Prediction

Reference number	Description	Results produced in the References
1	This paper has summarised state of art techniques and methods such as Decision Tree, Naive Bayes, ANN, SVM, and Linear Regression for prediction of heart disease.	SVM and Neural Network are considered as major competitive machine learning algorithms. SVM provides high accuracy to each dataset with high dimensionality.
2	Proposed comparative analysis between different classification algorithms, namely, J48, REPTREE, SIMPLE CART, Naive Bayes, Bayes Net using different evaluation measures such as Timing to build model, Correctly classified instances ,incorrectly classified instances, predictive accuracy, Kappa Statistics, Mean absolute Error, Root Mean Squared, Relative Absolute Error, and Root Relative Squared Error	The accuracy obtained with for different classifiers are given as: J48 - 99.0741% REPTREE - 99.07% SIMPLE CART - 99.0741% Naive Bayes - 97.222% Bayes Net - 98.148%.
3	Proposed comparative analysis between K- means based MAFIA, K-means based MAFIA with ID3, K means based MAFIA with ID3 and C4.5 based on Precision, Recall and Accuracy.	K-means based MAFIA with ID3 and C4.5 has given better precision, recall and accuracy compared to that produced by K-means based MAFIA with ID3 or with simple K-means based MAFIA.
4	Proposed comparative analysis between Neural Network, Naive Bayes, Decision Tree.	The accuracy obtained with for different classifiers are given as: Neural Network - 100% Naive Bayes - 90.74% Decision Tree - 99.62% Neural Network gives high accuracy than others
6	The objective is to predict the diagnosis of heart disease with reduced number of attributes by using Decision Tree, Naive Bayes, and Classification via clustering. In this work 14 attributes such as Age, Sex, CP, Rbp, Cholestrol, Fasting blood sugar, Resting ECG, Thalach, Induced angina, Old peak, Slope, Thal, CA are reduced into 6 attributes such as Rbp, Oldpk, Type(CP), CA(VSL), Eia(Exercise induced angina), Thal by using Genetic algorithm	The genetic search resulted in 6 attributes which contributes more towards the diagnosis of the cardiac disease. Decision Tree outperformed after reducing the attributes with high constructing time for model Naive Bayes performs consistently before and after attribute reduction Classification via Clustering performs poor compared to other two methods
7	This research has developed a prototype Intelligent Heart Disease Prediction System(IDHPS) using data mining techniques, namely Decision Tree, Naive Bayes, ANN.	Naive Bayes appears most effective as it has the highest percentage of correct predictions(86.53%) for patients with heart disease, followed by Neural Network with 85.53% Decision Tree is most effective in case of predicting patients with no heart disease(89%) . The analysis shows that Neural Network with 15 attributes has shown highest accuracy ie)100% and Decision Tree with 15 attributes accuracy is 99.62% In combination with genetic algorithm and 6 attributes, Decision Tree has shown 99.2%
8	Heart disease prediction system is developed using Neural Network. In this work 2 more attributes such as obesity and smoking are added along with 13 attributes such as age, sex, CP, thestbps, chol, restecg, FBS, thalach, exang, oldpeak, slope, ca and thal.	With 13 attributes, the prediction accuracy is 99.25% and with 15 attributes the accuracy is nearly 100%.
9	Survey tells the uses of different Decision tree algorithms for prediction of heart disease.	Recommends Decision Tree for prediction due its simplicity and availability of different attribute selection measures such as Information Gain, Gain Ratio and Gini Index
10	Proposed comparative analysis between Support Vector machine and Ensemble methods such as Bagging, Boosting, Random Subspace for heart disease prediction based on accuracy, sensitivity, specificity, PPV(Positive Prediction Value),NPV(Negative Prediction Value).	Bagging gives the accuracy of 81.35%, Boosting gives accuracy of is 83.22%, SVM gives accuracy is 73.7% and Random Subspace gives accuracy of 80.00%. So Boosting is found as better than others.
11	The main objective of this work is to provide a study of different data mining techniques that can be used in automating heart disease prediction system.	This paper reviews the literature and finds that SVM provides effective and efficient accuracy of 85% as compared to other data mining techniques.

12	Proposed heart disease prediction using Decision Tree and SVM classifiers. REPTREE is used for feature selection. Using 10 fold cross validation and four kernel types(Linear, Polynomial, sigmoid, RBF) and two SVM models(CSVM & NUSVM)	With 13 attributes the system gives an accuracy of 76.66%. After removing the attributes OP, RBF, Sex, SM , it gives an accuracy of 77.91%. On repeating adding and removing attributes accuracy percentage is increased into 82.15%. The number of attributes reduced into 3(Thal, CP, Cf) gives an accuracy of 88.16%. Sex and ECG does not improve the performance. Kernel function RBF for CSVM gives better accuracy.
13	This paper suggests the use of MapReduce framework over HDFS prediction of heart disease	Generally suggests the use of big data tools
5	This paper proposes the use of data mining techniques such as Naive Bayes and SVM on big data to predict heart attacks.	SVM gives highest performance of correctness followed by Naive Bayes.

III. INFERENCES FROM SURVERY

At first the survey paper are analysed for choosing attributes for heart disease prediction. From survey, it is found that all the mentioned references have used data set available from UCI repository. Heart disease data contains 76 attributes. [Please see Reference - 14]. From the mentioned references it is found that among these 76, thirteen attributes, namely, *age,sex,CP,thestbps,chol, restecg, FBS, thalach, exang, oldpeak, slope, ca and thal*. From these 13 attributes, in[6], genetic algorithm has been used to select 6 features, namely, Chest pain type, Rbps, exang, oldpeak, Thalach, Ca. Further from [6] it is understood that the 6 attributes resulted from genetic search contribute more towards the diagnosis of the cardiac disease. In [7-8], along with 13 attributes the authors have taken into account two more attributes, namely, smoke, obes for prediction of heart disease. The authors of [7-8] proved the prediction of heart disease with higher accuracy from 99.25% to near 100% by Neural Network.

Secondly from survey it is found that classification algorithms such as Decision Tree, Naive Bayes, Artificial Neural Network, Support Vector Machine are used for heart disease prediction. Among these Naive Bayes is a linear classifier which gives the accuracy of 90.74% with 15 attributes and 94.44% with 13 attributes and 96.53% with 6 attributes. The other mentioned algorithms are non linear classifiers. Decision Tree gives accuracy 99.62% with 15 attributes and 96.66% with 13 attributes. Artificial neural network gives 99.25% with 13 attributes. The above accuracy values are given in [9]. From [9], it is found that non linear classifier gives more accuracy than linear classifier

Thirdly, it is found that the research works have used tools such as WEKA, MongoDB, Orange, Matlab, Tangara.

IV. PROPOSED WORK

As it is inferred from[6] that 6 attributes namely *Chest pain type, Rbps, exang, oldpeak, Thalach, Ca* are sufficient enough to predict heart disease with required accuracy. Further, it is realized that the non linear classifier algorithms will give better accuracy for heart disease prediction than linear algorithms [9]. Hence in this present work, it is proposed to use non linear classification algorithm for heart disease prediction. In regard to non- linear classification algorithms, from [1, 10] it is clear that among various non linear classification algorithms such as Decision Tree, Neural Network, Support Vector Machine gives better accuracy. Support vector machine is widely accepted machine learning classifier algorithm because of its generalization capacity [13]. SVM provides better and efficient accuracy about 85% and 82.35% as mentioned in [11] and [12] respectively with their respective data sets. Also [5] shows that higher performance of correctness can be achieved with SVM.

Another important aspect while predicting prediction of heart disease is related to size or volume of data as with existing volume of data and the speed with which the data gets generated make the storage and processing as difficult task for the current computing infrastructure [13].

Keeping the above ideas of attributes, predictive algorithms, data set and tools, it is proposed to develop a framework which alleviates the difficulties associated with volume of data. It is proposed to use big data tools such as Hadoop Distributed File System (HDFS), Mapreduce along with SVM for prediction of heart disease with optimized attribute set.

1. It is proposed to take into account the 6 chief features namely *Chest pain type, Rbps, exang, oldpeak, Thalach, Ca* as these are proved to produce prediction with sufficient accuracy[6].
2. It is proposed to use SVM algorithm as it is reported as better one[1, 10, 11, 12, 13] for prediction when compared to other similar algorithms.
3. It is proposed to use data set UCI repository as it is the bench mark data set.

In order to handle large amount of heart disease data, it is proposed to store the data in different data nodes of HDFS. Further, it is proposed to perform the prediction process simultaneously in all data nodes the parallel programming paradigm, MapReduce technique is used. The block diagram of proposed concept is given in Fig. 1.

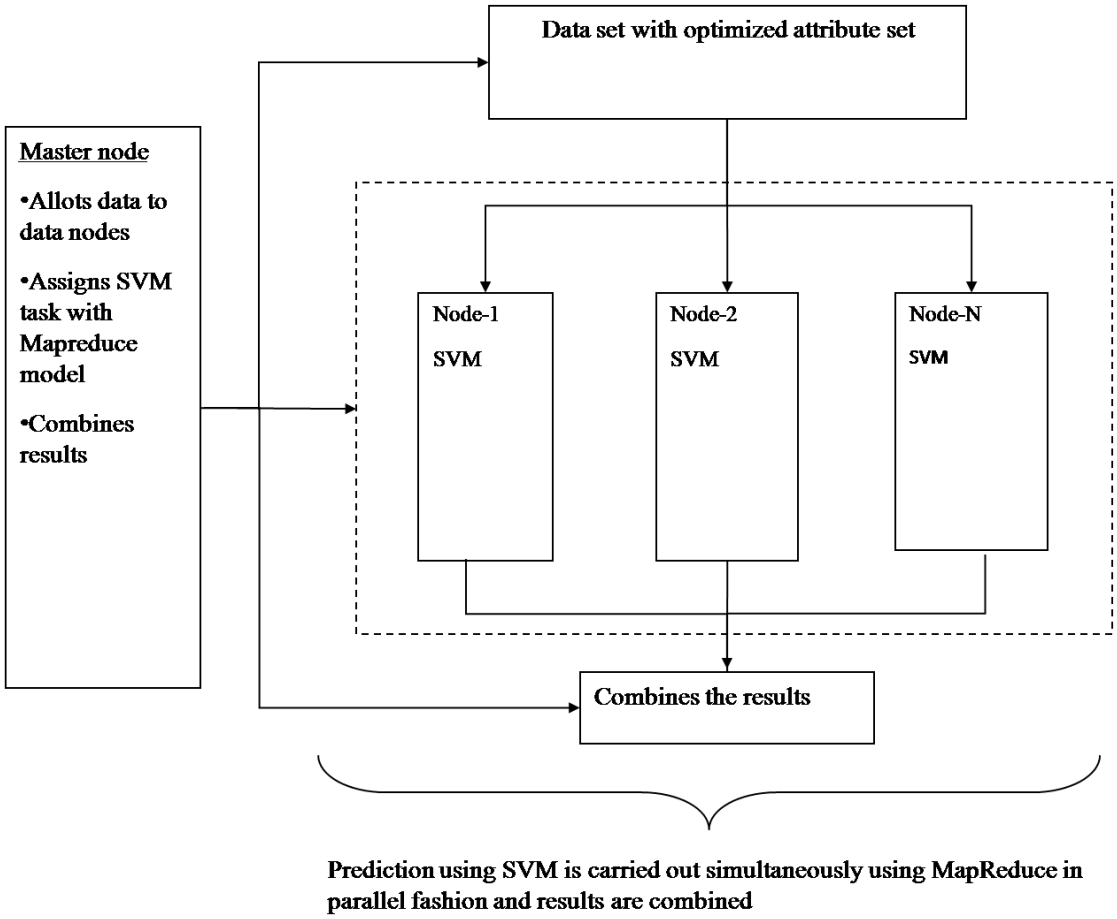

Figure 1. Block diagram of proposed approach

V. CONCLUSION

This work has made an investigation on the use of different data mining techniques for predicting heart diseases. It highlights the issue of handling large amount of data for prediction. It proposes an alternate method for predicting heart disease with large amount of data using big data techniques. It suggests the use of HDFS for storing large data in different nodes and executing the prediction algorithm using SVM in more than one node simultaneously using SVM. So, SVM is used in parallel fashion which will yield better computation time than sequential SVM.

REFERENCES

[1] Himanshu Sharma,M A Rizvi,"Prediction of Heart Disease using Machine Learning Algorithms:A survey",International Journal of Recent and Innovation Trends in Computing and Communication", Volume:**5**,Issue **8**,pp.**99-104**

[2] Himanshu Sharma,M A Rizvi,"Prediction of Heart Disease using Machine Learning Algorithms:A survey",International Journal of Recent and Innovation Trends in Computing and Communication", Volume:**5**,Issue **8**,pp.**99-104**

[3] V. Manikantan, S.Latha, "Predicting the Analysis of Heart Disease Symptoms using Medicinal Data Mining Methods, conference paper,**2013**,pp.**5-10**

[4] T.Revathi,S.Jeevitha,"Comparitive study on Heart Disease Prediction System using Data Mining Techniques", International Journal of Science and Research,**2013**,pp. **2120-2123**

[5] Pediredla Praveen Kumar,Sunita a Yadwad V V D L Tejaswi, "Prediction of Heart Disease using Hadoop Mapreduce", International Journal of Computer Application (2250-1797) Volume **6**– No.6, November – December **2016**. Pp.**1-8**

[6] Shamsher Bahadur Patel,Pramod Kumar Yadav,Dr.D.P.Shukla,"Predict the Diagnosis of Heart Disease Patients using classification Mining Techniques",Journal of Agricultural and Verterinary Science,Volume **4**,Issue **2**,**2013**,pp.**61-64**.

[7] Nidhi Bhatla,Kiran Jyoti,"An Analysis of Heart Disease Prediction Using Different Data Mining Techniques",International Journal Of Engineering Research & Technology,Volume1, Issue **8**,Oct-**2012**,pp.**1-4**

[8] Miss.Chaitrali.s,Dangare,Dr.Mrs.Sulabha S.Apte,"A Data Mining approach for prediction of heart disease using Neural Networks",International Journal of Computer Engineering & Technology,Volume **3**, October **2012**,pp.**30-40**

[9] K.Thenmozhi,"Heart Disease Prediction using Classification with Different Decision Tree Techniques",International Journal of Engineering Research and General Science ,Volume **2**,Issue **6**,October **2014**,pp.**6-11.**

[10] V. Subha and M.Revathi, D.Murugan," Comparative Analysis of Support Vector Machine Ensembles for the Heart disease prediction", International Journal of Computer Science & Communication Networks,Vol **5**(6),386 – 390,December **2015**,pp.**386-390**.

[11] Megha Shahi,Er.Rupinder Kaur Guram,"Heart Disease Prediction System Using Data Mining Techniques –A Review", International Journal of Technology and Computation, Volume **3**, Issue **4**, April **2017**,pp.**73-77**.

[12] Shalet K.S, V.Sabarinathan, V.Sugumaran, V.J.Sarath Kumar," Diagnosis of Heart disease using Decision Tree and SVM Classifier", International Journal of Applied Engineering Research,ISSN 0973-4562 Vol.**10** No.**68(2015)**,pp.**598-602**.

[13] Dr.Siddaraju,Sowmya.c l,Rashmi k,Rahul M,"Efficient Analysis of Big Data using Map Reduce Framework",International Journal of Recent Development in Engineering and Technology, Volume**2**,Issue **6**,June **2014**, pp.**64-68**

[14] https://archive.ics.uci.edu/ml/datasets/Heart+Disease**.**

Authors Profile

Mrs.R. Sharmila pursed Master of Computer Applications from Bharathidasan University of India in 2002. She is currently pursuing Ph.D. and currently working as Lecturer in Department of Computational Sciences, Bharathidasan University Constituent Arts & Science College, Trichy, India. Her main research work focuses on Big Data Analytics and Data Mining. She has 12 years of teaching experience and 3 years of Research Experience.

Dr. S. Chellammal has 10 years of Industry and R&D experience along with 8 years of academic experience. She has been a life member in CSI and IAENG. She has published papers in IEEE Transactions on Services Computing and Springer SOCA. She published book chpaters with IGI Global. Her research work focuses on semantic web services, data mining, text analytics and big data.

