International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

An Enhancing Heart Disease Prediction Using Machine Learning Algorithms

Sudha.K[1], Anupriya.K[2] , Banumathi.C[3], Dr.P.Srinivasan[4]

Associate Professor, Department of Computer Engineering, Muthayammal Engineering College, Rasipuram, Tamilnadu, India[1]

Professor, Department of Computer Engineering, Muthayammal Engineering College, Rasipuram, Tamilnadu, India[4]

U.G. Student, Department of Computer Engineering , Muthayammal Engineering College, Rasipuram, Tamilnadu, India [2&3]

ABSTRACT*:* Heart disease is considered as one of the major causes of death through out the world. It cannot be easily predicted by the medical practitioners as it is a difficult task which demands on expertise and higher knowledge for the prediction. An automated system in medical diagnosis would enhance medical efficiency and also reduce cost. The goal is to pick hidden patterns by applying data processing techniques, which are note worthy to heart diseases and to predict the presence of heart disease in patients where the presence is valued on a scale. Our objective is to find out the suitable machine learning technique such as that is computationally efficient as well as accurate for the prediction of heart disease. So, this research focuses on feature selection techniques and algorithms where multiple heart disease datasets are used for experimentation analysis and to show the accuracy improvement. By using the Rapid miner as tool; Logistic Regression, Logistic Regression SVM, Naïve Bayes and Random Forest; algorithms are used as feature selection techniques and improvement is shown in the results by showing the accuracy.

KEYWORDS: Naive Bayes, Logistic Regression, PCA (Personal Component Analysis), SVM (Support Vector Machine).

I . INTRODUCTION

The contents of this paper mainly focus on various data mining practices that are valuable in heart disease forecast with the assistance of dissimilar data mining tools that are accessible. If the heart doesn't function properly, this will distress the other parts of the human body such as brain, kidney etc. Heart disease is a kind of disease which effects the functioning of the heart. In today's era heart disease is the primary reason for deaths. WHO-World Health Organization has anticipated that 12 million people die every year because of heart diseases. Some heart diseases are cardiovascular, heart attack, coronary and knock. Knock is a sort of heart disease that occurs due to strengthening, blocking or lessening of blood vessels which drive through the brain or it can also be initiated by high blood pressure.

The major challenge that the Healthcare industry faces now-a-days is superiority of facility. Diagnosing the disease correctly & providing effective treatment to patients will define the quality of service. Poor diagnosis causes disastrous consequences that are not accepted. Records or data of medical history is very large, but these are from many dissimilar foundations. The interpretations that are done by physicians are essential components of these data. The data in real world might be noisy, incomplete and inconsistent, so data preprocessing will be required in directive to fill the omitted values in the database. Even if cardiovascular diseases is found as the important source of death in world in ancient years, these have been announced as the most avoidable and manageable diseases. The whole and accurate management of a disease rest on the well-timed judgment of that disease.

Heart expert's create a good and huge record of patient's database and store them. It also delivers a great prospect for mining a valued knowledge from such sort of datasets. There is huge research going on to determine heart disease risk factors in different patients, different researchers are using various statistical approaches and numerous programs of data mining approaches. Statistical analysis have acknowledged the count of risk factors for heart diseases counting smoking, age, blood pressure, diabetes, total cholesterol, and hypertension, heart disease training in family, obesity and lack of exercise. For prevention and healthcare of patients who are about to have addicted of heart disease it is very important to have awareness of heart diseases. Researchers make use of several data mining techniques that are accessible to help the specialists or physicians identify the heart disease. Commonly used procedures used are decision tree, k-nearest and Naïve Bayes. Other different classification based techniques used are bagging algorithm, kernel

density, sequential minimal optimization and neural networks, straight Kernel self-organizing map and SVM (Support Vector Machine). The next section clearly provides details of techniques that were used in the study.

II . RELATED WORK

Authors of [1] had used the healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining technique scan help remedy this situation.research has developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naive Bayes and Neural Network.

[2] has been used an Android based prototype software has been developed by integrating clinical data obtained from patients admitted with IHD (Ischemic Heart Disease. The clinical data from 787patients has been analyzed and correlated with the risk factors like Hypertension, Diabetes, Dyslipidemia (Abnormal cholesterol), Smoking, Family History, Obesity, Stress and existing clinical symptom which may suggest underlying non detected IHD. The data was mined with data mining technology and a score is generated. Risks are classified into low, medium and high for IHD.

Authors of [3] had used heart disease is considered as one of the major causes of death throughout the world. It cannot be easily predicted by the medical practitioners as it is a difficult task which demands expertise and higher knowledge for prediction. This paper addresses the issue of prediction of heart disease according to input attributes on the basis of data mining techniques. We have investigated the heart disease prediction using KStar, J48, SMO, Bayes Net and Multilayer Perceptron through Weka software. The performance of these data mining techniques is measured by combining the results of predictive accuracy, ROC curve and AUC value using a standard data set as well as a collected data set. Based on performance factor SMO and Bayes Net techniques show optimum performances than the performances of KStar, Multilayer Perceptron and J48 techniques.

Authors of [4] had used Coronary artery disease is the leading cause of death in the world. In this research, we propose an algorithm based on the machine learning techniques to predict the risk of coronary artery atherosclerosis. A ridge expectation maximization imputation (REMI) technique is proposed to estimate the missing values in the atherosclerosis databases. A conditional likelihood maximization method is used to remove irrelevant attributes and reduce the size of feature space and thus improve the speed of the learning. The STULONG and UCI databases are used to evaluate the proposed algorithm.

The performance of heart disease prediction for two classification models is analyzed and compared to previous work. Experimental results show the improved accuracy percentage of risk prediction of our proposed method compared to other works. The effect of missing value imputation on the prediction performance is also evaluated and the proposed REMI approach performs significantly better than conventional techniques.

Authors of [5] had used We are in a period of "Information Age" where the traditional industry can pressure the rapid shift to the industrial revolution for industrialization, based on economy of information technology Terabytes of data are produced and stored day-to-day life because of fast growth in "Information Technology". Terabytes of data are produced and stored day-to-day life because of fast growth in "Information Technology". The data which is collected is converted into knowledge by data analysis by using various combinations of algorithms.

For example: the huge amount of the data regarding the patients is generated by the hospitals such as x-ray results , lungs results ,heart paining results, chest pain results , personal health records(PHRs)., etc. There is no effective use of the data which is generated from the hospitals. Authors of [6] had used the main theme of the paper is the prediction of heart diseases using machine learning techniques by summarizing the few current researches. In this paper the logistic regression algorithms is used and the health care data which classifies the patients whether they are having heart diseases or not according to the information in the record. Also I will try to use this data a model which predicts the patient whether they are having heart disease or not.

III . METHODOLOGY

We propose 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. The prediction model is introduced with different combinations of features and several known classification techniques. The Logistic Regression with SVM known as Logistic Regression SVM (LRSVM) is introduced for the diagnosis of heart disease. Heart disease prediction with Logistic Regression, Logistic Regression SVM, Naive Bayes and Random Forest is proposed. In this approach, four methods are used for the premise of the accuracy and time of testing.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

The proposed model arranges the data records into two classes for further analysis. A Data mining model has been developed using Random Forest classifier to improve the prediction accuracy and to investigate various events related to CHD. This model can help the medical practitioners for predicting CHD with its various events and how it might be related with different segments of the population.

A. Classification Algorithms

- Logistic regression: Logistic regression is assessing the parameters of logistic model in regression analysis.

- PCA: This algorithm is an unsupervised, non-parametric statistical technique primarily used for dimensionality reduction in ML.PCA can also be used reduce number of variables in your data by extracting important one from a large pool.

- Random forest:Random forest is a tree based method that is used for both classification and regression analysis. Multiple trees are constructed and the mean prediction would be the output for classification.

- Naive Bayes:Naive Bayes classifiers perform classification by calculating the probability of given dataset. Each attribute in given data is considered as independent of other.

B. Proposed Approach

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 pre-processing.

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 value is set to 1, else the value is set to 0 indicating the absence of heart disease in the patient. 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.

From among the 14 attributes of the data set, two attributes pertaining to age and sex are used to identify the personal information of the patient. The remaining 7attributes 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, LR, RF,PCA and SVM. The experiment was repeated with all the ML techniques using all 14 attributes in prediction method of PCA.

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 In Fig1.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

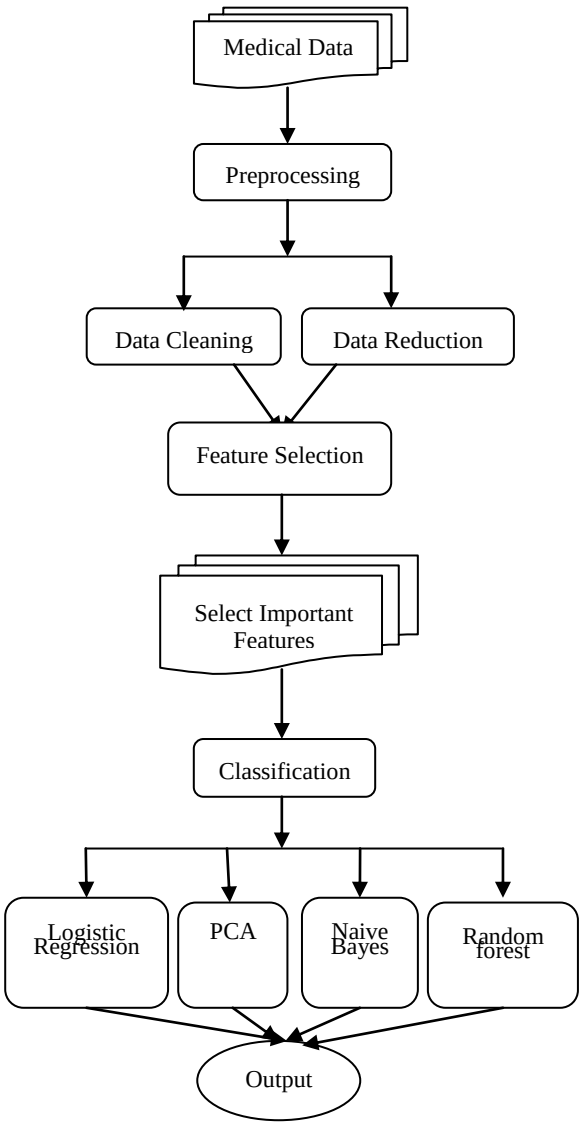

Fig1:Proposed System Approach

C. Performance Analysis

Though the final result is generated from only a single algorithm, we test four different algorithms in the model to compare their accuracies. These classification algorithms are Naïve Bayes, Logistic Regression, Random Forest, PCA Algorithm.In the proposed system, the PCA algorithm is the most efficient algorithm with an accuracy of 98%, with Naïve Bayes algorithm being the least efficient with an accuracy of 78%.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

Fig2: A sample dataset Analysis

S.No	Techniques	Data set	Accuracy
1	Logistic Regression	UCI	88%
2	PCA	UCI	98%
3	Navie Bayes	UCI	78%
4	Random Forest	UCI	92%

TABLE1: *Comparison of accuracy results*

IV. CONCLUSION&FUTURE WORK

The accuracy of Logistic Regression 88%, Random Forest 92%, Naive Bayes 78% and PCA is 97%. As if we compare this increase with the previous used techniques then the highest accuracy achieved was 97%. The PCA technique increase accuracy than previously implemented machine learning techniques. As the PCA is with higher accuracy achievement so this paper suggests PCA as a best feature selection technique for predicting heart disease.If we compare these techniques with rule mining which is also used previously with achievement of 97% accuracy, the proposed research improved much higher accuracy than rule mining.

Today's, world most of the data is computerized, the data is distributed and it is not utilizing properly. By Analyzing the available data we can also use for unknown patterns. The primary motive of this research is the prediction of heart diseases with high rate of accuracy. For predicting the heart disease we can use Random Forest, Logistic Regression algorithm, Naive Bayes, PCA in machine learning. The future scope of the paper is the prediction of heart diseases by using advanced techniques and algorithms in less time complexity. In future these techniques can also be applied on real time medical datasets and also can be used in the form of ensembles i.e. combinations of multiple techniques. This would result in increase of further accuracy and high performance.

REFERENCES

[1] SellappanPalaniappan, RafiahAwang, "Intelligent heart disease prediction system using data mining techniques", 2008 IEEE/ACS International Conference on Computer Systems and Applications, 22 April 2008.

[2] M. Raihan, SaikatMondal, Arun More, "Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design", 2016 19th International Conference on Computer and Information Technology (ICCIT), 23 February 2017.

[3] Marjia Sultana, Afrin Haider , "Analysis of data mining techniques for heart disease prediction", 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 09 March 2017.
[4] SoodehNikan, FemidaGwadry-Sridhar, "Machine Learning Application to Predict the Risk of Coronary Artery Atherosclerosis", 2016 International Conference on Computational Science and Computational Intelligence (CSCI), 20 March 2017.
[5] Reddy Prasad, Pidaparthi Anjali, S.Adil, N.Deepa, "Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning" International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8, Issue-3S, February 2019.
[6] ARCHANA, BADE, AHER DIPALI, and SMITA KULKARNI PROF. "International Journal On Recent and Innovation Trends In Computing and Communication." 2277-4804.
[7] Silwattananusarn, Tipawan, and KulthidaTuamsuk. "Data mining and its applications for knowledge management: A literature review from 2007 to 2012." arXiv preprint arXiv:1210.2872 (2012).
[8] K.Sudha.," An Efficient Two-Factor Access Control for Web-based Cloud Computing Services using Jar File" International Journal of Scientific Research in Computer Science,Engineering and Information Technology Vol 2, Issues 2, April 2017.
[9] K. Sudha "Load balancing and Monitoring System using New Cracking Algorithm" International Journal of Advanced Research in Biology Engineering Science and Technology Vol 2, Issues 2, Feb 2016
[10] Leventhal, Barry. "An introduction to data mining and other techniques for advanced analytics." Journal of Direct, Data and Digital Marketing Practice 12, no. 2 (2010): 137-153.
[11] McKhann, Guy, David Drachman, Marshall Folstein, Robert Katzman, Donald Price, and Emanuel M. Stadlan. "Clinical diagnosis of Alzheimer's disease Report of the NINCDS- ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease." Neurology 34, no. 7 (1984): 939-939.
[12] Nahar, Jesmin, Tasadduq Imam, Kevin S. Tickle, and Yi-Ping Phoebe Chen. "Association rule mining to detect factors which contribute to heart disease in males and females." Expert Systems with Applications 40, no. 4 (2013): 1086-1093.
[13] Prakash, S., K. Sangeetha, and N. Ramkumar. "An optimal criterion feature selection method for prediction and effective analysis of heart disease." Cluster Computing (2018): 1-7.
[14] P.Srinivasan,M.Menakapriya and P.Suresh"Web based unique link recommendations using Fuzzy weight ranking algorithms", TAGAJOURNAL, SWANSEA PRINTING TECHNOLOGY LTD, ISSN: 1748-0345, Vol. 14, pp: 3200 – 3209, 2018.
[15] P.Srinivasan,M.Menakapriya and P.Suresh"Secured code construction for Information transfer between Smart phones", TAGAJOURNAL, SWANSEA PRINTING TECHNOLOGY LTD, ISSN: 1748-0345, Vol. 14, pp: 3210 – 3218, 2018.
[16] Sudha.K, "Machine Learning Approach on Embodied Intelligence" International Journal of Innovative Research in Engineering Science and Technology Vol VI, Issues 2, April 2018.
[17] Sudha.K, "identifying cardiovascular and eye disease using retinal vessel and OD "segmentation in international journal of research in computer application and robotics, volume to issue 5, ISSN:2320-7345 May 14.
[18] Sudha.K, published a paper on "Finding an optimal forest graph to identify the cardio vascular diseases" Interantional Journal of Advanced and Innovative Research(IJAIR), volume2,issue 12, ISSN: 2278-7844, December 2013.
[19] Pouriyeh, Seyedamin, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, and Juan Gutierrez. "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease." In Computers and Communications (ISCC), 2017 IEEE Symposium on,pp.204-207.IEEE,2017