IJCSEInternational Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-8, Issue-2, Feb 2020 E-ISSN: 2347-2693

A Survey on Early Detection and Prediction of Heart Diseases using Machine Learning and Data Mining Techniques

Bhumika J.[1*], Rashmi R. Kotiyan[2], Sonal T.H.[3], Lakshmi R.[4]

[1,2,3,4]Information Science And Engineering, Global Academy of Technology, VTU, Bangalore, Karnataka, India

[*]*Corresponding Author: bhumikaaradhya98@gmail.com, Tel.: +91 7349790278*

DOI: https://doi.org/10.26438/ijcse/v8i2.3134 | Available online at: www.ijcseonline.org

Accepted: 10/Feb/2020, Published: 28/Feb/2020

Abstract— Cardio vascular disease is the most prominent cause of death worldwide. Machine Learning Algorithms can be used for predicting chances of heart disease occurrence. Relating machine learning and data mining methods is a strategic approach to consume large volumes of available Cardio-related data for prediction. The datasets used are classified in terms of medical parameters. In this paper, numerous algorithms and techniques are discussed that are used in prediction of Cardio Vascular Diseases. Fast Correlation-Based Feature Selection (FCBF) method to filter noise data to improve quality of heart disease classification. K-Nearest Neighbour, Support Vector Machine, Naïve Bayes, Random Forest and a Multilayer Perception, Artificial Neural Network optimized by Particle Swarm Optimization (PSO) combined with Ant Colony Optimization (ACO) are the classification algorithms used. By using machine learning algorithms and deep learning it provides numerous ways for the prediction of the heart disease. There are various methods which provide us an information and these are applied to various datasets to get particular results.

Keywords— Heart Disease, Predictive Analysis, Naïve Bayes, Decision Tree, SVM

I. INTRODUCTION

As heart is an important organ in human body, if heart doesn't work properly it will affect the whole human body [2]. Even though there are various instruments available, they are very expensive to use and there is no guarantee that they would predict the heart disease accurately [5]. So Machine Learning is the best choice for predicting a heart disease. 17 million people die every year because of heart diseases like heart attack and stroke as estimated. It is important to record the symptoms, health and food habits that are contributing to heart diseases [1]. It is even worse in some of the developing countries where the lifestyle of people is very poor. High Blood Pressure, obesity and smoking are some of the factors that turns out to be a major affecting parameter. It has even been difficult for physicians to find out the reasons for the particular heart disease. It is also seen that there is a shortage of apparatus which fails to predict the disease at early stage [6]. Hence we need to find a suitable approach which would control the death rate. Various techniques proposed in data mining and neural networks have been used to find out the particular heart disease in humans[3]. Machine Learning uses vector, data types and several other methods like K-Nearest Neighbor Algorithm(KNN), Decision Trees(DT), Genetic Algorithm (GA), and Naïve Bayes(NB) which reduces the error in the prediction[5]. These Algorithms are very useful for this generation and they are used more frequently in medical organisations [7].These techniques saves million lives and also reduces the workload of doctors [7]. Due to this death rate has been reduced compared to the previous times [6].

II. RELATED WORK

[1]Youness Khourdifi, Mohamed Bahaj, focused on optimization algorithms since it has the advantage of dealing with complex non-linear problems. The classification algorithms such as K-Nearest Neighbor, Support vector machine, Naïve Bayes, Random Forest and Multilayer perception, Artificial neural network is been implemented. These proposed mixed approaches is applied to heart disease datasets.

[2] S. Kavitha, K.R. Baskaran, S. Sathyavathi, mainly focused on using Naïve mathematician techniques to predict the heart disease. Neural network provides reduced error for the prediction of heart disease.

[3] SenthilKumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava, have mainly focused on using Data Mining techniques to discover various sorts of metabolic syndromes. Decision trees are also used for predicting the accuracy of events in heart disease. Several standard performance metrics such as accuracy, precision and error in

classification has been considered for the computation of performance efficiency. Highest accuracy has been achieved by HRFLM classification method.

[4] Niraj Kalantri, Kumar R, have focused on using data mining techniques. The major reason of using it was the availability of the data and turning the same data into some useful information. Classification algorithms like Decision Tree, Random Forest, SVM, Logistic regression and K-Nearest neighbour are implemented on training data set for the accurate prediction of heart disease. After the classification, the highest accuracy was predicted using Logistic Regression.

[5] Amin Ul Haq, Jian Pinf Li, Muhammed Hammad emon, Shah Nazir and Rulnan Sun., have made use of seven popular algorithms like three feature selection algorithms, the cross-validation method, and seven classifiers performance evaluation metrics such as classification accuracy, specificity, sensitivity, Matthews correlation coefficient and execution time. These algorithms provides best Accuracy.
[6] Himanshu Sharma M A Rizvi made use of decision tree, Support vector machine, deep learning, k-nearest neighbor algorithms. Since datasets contain noise, thus noise were reduced by cleaning and pre-processing the datasets.

[7] V.V.Ramalingam, Ayantan Dandapath, M Karthik Raja have used Models based on supervised learning algorithms such as Support Vector Machines(SVM), K-Nearest-Neighbour(KNN), Naive Bayes, Decision Trees(DT), RandomForest(RT), and ensemble models are found very popular among the researchers.

III. METHODOLOGY

Heart disease datasets are collected from UCI repository. Noise data are eliminated by pre-processing and selects the best features [2]. Machine Learning process starts from pre-processing data, followed by feature selection, classification of modelling, performance evaluation and finally the results. Feature combination and modelling keeps on repeating for various combinations of attributes [4]. Some of the proposed methods like K-Nearest Neighbour, Support Vector Machine, Naïve Bayes, Random Forest and Artificial Neural Network are used for classification[1]. The K-Nearest Algorithm is an effective classification technique [7].The decision Tree approach is very understandable approach, in this each leaf is treated as the node [6].The SVM is mainly used for the classification problems and it also helps to solve the complex problems[6].The Naive Bayes uses the training datasets to find the probability of the given classes.

3.1 Work Flow

3.1.1 Selection
Selection of an appropriate datasets is needed for the accurate prediction of heart diseases. Sufficient quantity of data is needed to perform Machine Learning techniques on selected heart disease dataset.

3.1.2 Pre-processing and Transformation
CSV (Comma Separated Values) file format is prepared for the dataset. Data pre-processing is already done with "?" in the place of missing values. Normalization has been done as a part of transformation on the data.

3.1.3 Training and Testing
Here only 20% of the data is kept for testing and remaining all is sent for training the data. After the data is trained, classification is done on testing data in order to get the result.

3.1.4 Applying Appropriate Classification Algorithm
The classification algorithms like Decision Tree, Random Forest, SVM, Logistic regression and k-nearest neighbour are implemented on trained dataset [4].

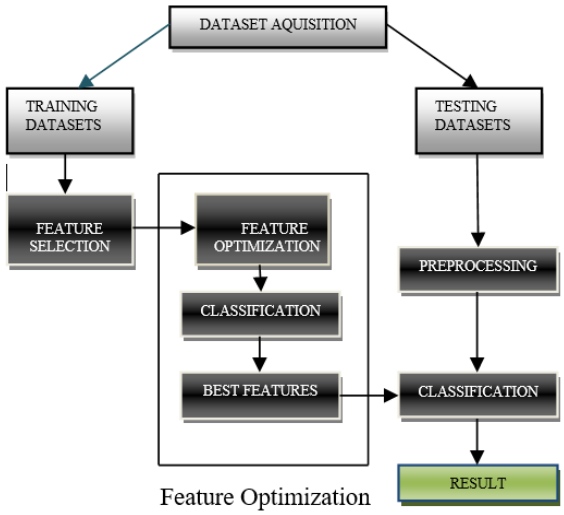

Figure 1. The Proposed Architecture

IV. RESULTS AND DISCUSSION

The K-Nearest Algorithm is an effective classification technique [7]. It is used where there is no prior knowledge about data distribution. This algorithm involves finding k nearest points in training sets to the point in which target value is unavailable. It is better than other technique with less error rate.

The decision Tree approach is very understandable approach. In this each leaf is treated as the node [6]. It is mainly used for classification problems. Based on the significant predictors this algorithm divides the population into two or more similar sets. This algorithm first calculates the entropy of each attributes. Using the maximum entropy gain the dataset is spitted with the help of variables. Finally these same steps are repeated for remaining attributes.

The SVM is mainly used for the classification problems and it also helps to solve the complex problems [6]. It can be used as classifier as well as predictor. It finds the hyper-plane that differentiates between the classes [7]. The training data points in feature space are mapped in such a way that points that belongs to separate classes are segregated by a margin as wide as possible and test data points are classified based on the side which the margin fall.

The Naive Bayes is an effective technique which is based on the Bayes theorem. It uses the training datasets to find the probability of the given classes. In this the attributes can be related to each other.

Random Forest can be used for both regression and classification tasks which performs better. Before giving the output it considers multiple decision trees. It decides the class through the voting where the regression takes the mean of all outputs of each decision tree and also works with large datasets[7].

The main aim was to test which algorithm classifies the heart disease with the proposed optimization methods. 10-fold cross validation is used due to small number of collected features. Each experiment is run several times to avoid the instable operation result and for the comparison, classification accuracy was selected[2]. Classification models and their outcomes are also involved. Performance of different algorithms such as K-Nearest, artificial neural network, Naïve Bayes and decision tree are used using Cleveland heart disease datasets on features. For important features selection, feature selection algorithms are used. Performances were checked using the classifiers for selected features. Evaluation metrics were applied to check the performance of classifiers. Before applying to the classifiers features were normalized [4].

V. CONCLUSION

Identifying and processing the raw data of heart information will help in early detection of heart disease and reducing the number of mortality rates in the world. Machine Learning algorithms and techniques are used for processing the data. Heart disease prediction has been a challenging one in the medical field. From the experimental results, the execution time calculated for classification object is sort of reduced than the present system. Each algorithm worked better than previously existed system and they were very fast and performed very well. The important algorithms are used to select the important features. For classifiers, different evaluation metrics are taken into consideration. The feature selection algorithms reduced the execution time of the system and these systems increased the performance.

VI. FUTURE SCOPE

There are several algorithms that are been used for predicting the accuracy of heart disease. Moreover, it has been noticed that accuracy of the same algorithms differ according to the way of training them. Hence the future work will be to increase the accuracy of these algorithms. The work will also be on focusing to integrate different datasets and so that heart disease can be predicted with more accuracy.

ACKNOWLEDGMENT

I would like to express my heartfelt gratitude and sincere thanks to my faculty Mrs. Lakshmi R for her support and guidance throughout the course of my research process. Her valuable comments and sharing of knowledge of related work are the added advantage for the research work. Any of the errors in this project are on my own and should not tarnish the reputation of the esteemed person.

REFERENCES

[1]. Youness Khourdifi Mohamed Bahaj, ”Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization”, International Journal Of Intelligent Engineering and System, **October-2018**.

[2]. S.Kavitha, K.R.Baskaran,S.Sathyavathi, ”Heart Disease with Risk Prediction using Machine Learning Algorithms”, International Journal of Recent Technology and Engineering (IJRTE), ISSN: **2277-3878**, Volume**:7** ,Issue:**4S**, pp. **314-317**, **November 2018**.

[3]. SenthilKumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”, IEEE Access, ISSN:**2923-707**, Volume:**10**, **July 3 2019.**

[4]. Niraj Kalantri, Kumar R, “Predictive Analysis on Heart Disease Using Different Machine Learning Techniques”, Internation Journal of Computer Science And Engineering, ISSN:**97-101**, Volume:**7**, Issue:**2**, **28-Feb-2019**.

[5]. Amin Ul Haq , Jian Ping Li ,Muhammad Hammad Memon ,Shah Nazir ,and Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", Hindawi Mobile Information Systems, ISSN:**3860-146**, Volume: **2018**, pp. **1-21**

[6]. Himanshu Sharma,M A Rizvi,"Prediction of Heart Disease using Machine Learning Algorithms", International Journal On Recent And Innovative Trends In Computing And Communication, ISSN: **2321-8169**, Volume:**5**, Issue:**8**, pp. **99-104**, **August-2017**

[7]. V.V. Ramalingam, Ayantan Dandapath, M Karthik Raja, "Heart disease prediction using machine learning techniques", International Journal Of Engineering And Technology, Volume:**7**, Issue:**2.5**, pp. **684-687, March 2018**

[8]. Apurva Gaikwad, M.S. Panse, “Extraction of FECG from Non-Invasive AECG signal for Fetal Heart Rate Calculation”, ISSN:**2321-3256**, Volume:**5**, Issue:**3**, pp. **1069-112**, **2017**

[9]. Alister Dsouza, M. S. Panse, “LabVIEW based detection of Pulse Transit Time from Plethysmogram and ECG signals for estimation of Blood Pressure”, ISSN: **2320-7639**, Volume:**5**, Issue:**4**, pp**. 36-40**, **2017**

Authors Profile

Ms.Bhumika J pursuing BE in Global Academy of Technology in the Department of Information Science and Engineering Bangalore . Her areas of interest are Machine Learning,Big Data and Computer Networking.

Ms.Rashmi R Kotiyan pursuing BE in Global Academy of Technology in the Department of Information Science and Engineering Bangalore . Her areas of interest are Machine Learning, MERN Stack development and Computer Networking.

Ms.Sonal TH pursuing BE in Global Academy of Technology in the Department of Information Science and Engineering Bangalore . Her areas of interest are GUI development, Machine Learning, MERN Stack development, Cyber Security and Computer Networking.

Prof. Lakshmi R pursed B.E in Information Science and Engineering and M.Tech in Computer Network Engineering. She is currently working as a Professor in the Department of Information Science and Engineering. Her areas of interest are Network Security, Cybersecurity, and Cryptography.

