Prediction and Diagnosis of Cardiovascular Diseases Using Machine Learning: A Review

Shaswat Babhulgaonkar[1], Jayesh Suryavanshi[2], Pritam Bendkule[3], Professor L.A.Patil[4]
[1,2,3] *Student, Computer Dept., K.K.W.I.E.E.R, Savitribai Phule Pune University, India*
[4]*Asst. Prof, Computer Dept., K.K.W.I.E.E.R, Savitribai Phule Pune University, India*

***Abstract*- Diagnosis and Prediction of cardiovascular diseases has often become a challenge faced by doctors and hospitals in India as well as abroad. Despite major transformations in lifestyles of people and advancements in medical domain; heart attacks still hold a major share in the global death rate. The ambiguity in diagnosis of most heart diseases lies in the intricate grouping of clinical and pathological data which may introduce misinterpretation of data among clinical experts, doctors and researchers. Ultimately, the problem lies within making decisions concerned with predicting and later diagnosing the heart diseases. These decisions can have a drastic effect on life of a person. The proposed approach to use machine learning for prediction as well as diagnostic purposes can play a very important role in this area. Various Machine Learning techniques can be used for classifying healthy people from the ones suffering from heart diseases. This work intends to present a comprehensive review of prediction of Cardiac diseases by using Machine Learning based approach.**

Index Terms- Cardiovascular diseases, Data Mining, Machine Learning, Neural Networks, SVM

I. INTRODUCTION

Cardiovascular diseases is considered as one of the major causes of death in the medical field. It cannot be easily predicted by the medical practitioners as it is a difficult task which demands expertise and higher knowledge for prediction. An automated system in medical diagnosis would enhance medical efficiency to discover the rules that predict the risk level of patients based on the given parameters about their health. The goal 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 the presence is valued on a scale.

The prediction of heart disease requires a huge size of data which is too complex and massive to process and analyze by conventional techniques. Our objective is to find out the suitable machine learning technique that is computationally efficient as well as accurate for the prediction of heart disease. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. Common risk factors of heart disease include high blood pressure, abnormal blood lipids, use of tobacco, obesity, physical inactivity, diabetes, age, etc. which needs to be reduced to avoid heart diseases.

Data mining is the process of automatically extracting knowledgeable information from huge amounts of data. It has become increasingly important as real life data enormously increasing. Heart disease prediction system can assist medical professionals in predicting state of heart, based on the clinical data of patients fed into the system. There are many tools available which use prediction algorithms but they have some flaws. Most of the tools cannot handle big data. There are many hospitals and healthcare industries which collect huge amounts of patient data which becomes difficult to handle with currently existing systems. Machine learning algorithm plays a vital role in analyzing and deriving hidden knowledge and information from these data sets and improves accuracy and speed.[1]

II. LITERATURE SURVEY

A. Intelligent Heart Disease Prediction System Using Data Mining Techniques

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 techniques can help remedy this situation. This research has developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely decision Trees, naive bayes and neural network.[2]

B. Smartphone based Ischemic Heart Disease (Heart Attack) Risk Prediction using Clinical Data and Data Mining Approaches, a Prototype Design 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 787 patients 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.[1]

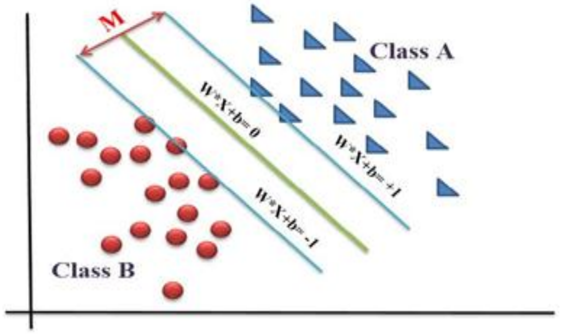
C. Analysis of Data Mining Techniques for Heart Disease Prediction Heart disease is considered as one of the major causes of death throughout the world. It cannot be easily predicted by the medical practitioners as it is a difficult task which demands expertise and higher knowledge for prediction. It 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 K Star, 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 K Star, Multilayer Perceptron and J48 techniques.[2]

D. Machine Learning Application to predict risk of Coronary Artery Atherosclerosis Coronary artery disease is the leading cause of death in the world. In this research, an algorithm is proposed 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.[3]

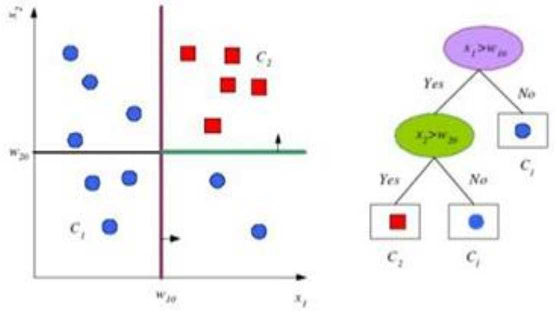
E. Prediction of Heart Disease Using Neural Network Heart disease is a deadly disease that large population of people around the world suffers from. When considering death rates and large number of people who suffers from heart disease, it is revealed how important is early diagnosis of heart disease. Traditional way of diagnosis is not sufficient for such an illness. Developing a medical diagnosis system based on machine learning for prediction of heart disease provides more accurate diagnosis than traditional way. In this a heart disease prediction system uses artificial neural network backpropagation algorithm. The thirteen clinical features were used as an input for the neural network and then the neural network was trained with backpropagation algorithm to predict the absence or presence of heart disease with good accuracy.[4]

III. PROPOSED TECHNIQUES

A. Support Vector Machine: Support vector machine (SVM) is supervised learning method that analyze data used for classification and regression analysis. It is given a set of training data, marked as belonging to either one of two categories. An SVM training algorithm then builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. The points are separated based on hyper planes that separate them. When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups.[3]



B. Decision Tree: A decision tree is a flowchart-like structure in which each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules. In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated. A decision tree consists of three types of nodes: Decision nodes represented by squares chance nodes, circles end nodes and triangles.Decision trees are commonly used in operations research and operations management. If in practice, decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Another use of decision trees is as a descriptive means for calculating conditional probabilities. Decision trees, influence diagrams, utility functions, and other decision analysis tools are taught to undergraduate students in schools of business, health economics, and public health, and are examples of operations research or management science methods.[5]

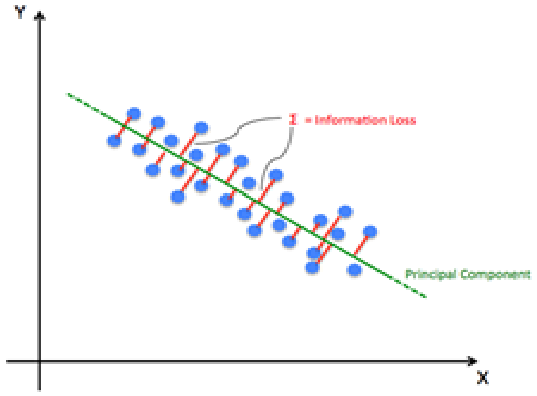


C. Naive-Bayes: In machine learning we are often interested in selecting the best hypothesis (h) given data (d).In a classification problem, our hypothesis (h) may be the class to assign for a new data instance (d).One of the easiest ways of selecting the most probable hypothesis given the data that we have which we can use as our prior knowledge about the problem. Bayes Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge. In Natural language processing a Naive Bayes Classifier is used to categorize the input data. It works on the basis of posterior probability which is computed using prior probability.[6]

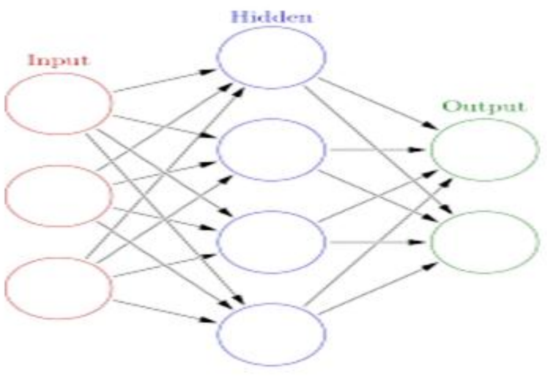
$$\underset{\text{Posterior Probability}}{P(c\mid x)}=\frac{\overset{\text{Likelihood}}{P(x\mid c)}\,\overset{\text{Class Prior Probability}}{P(c)}}{\underset{\text{Predictor Prior Probability}}{P(x)}}$$

$$P(c\mid X)=P(x_1\mid c)\times P(x_2\mid c)\times\cdots\times P(x_n\mid c)\times P(c)$$

D. Principal Component Analysis: The main idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent. The same is done by transforming the variables to a new set of variables, which are known as the principal components (or simply, the PCs) and are orthogonal, ordered such that the retention of variation present in the original variables decreases as we move down in the order. So, in this way, the 1st principal component retains maximum variation that was present in the original components. The principal components are the eigenvectors of a covariance matrix, and hence they are orthogonal. Intuitively, Principal Component Analysis can supply the user with a lower-dimensional picture, a projection or "shadow" of this object when viewed from its most informative viewpoint.[3]



E. Neural Networks: An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal from one artificial neuron to another. An artificial neuron that receives a signal can process it and then signal additional artificial neurons connected to it. Typically, artificial neurons are aggregated into layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times. The original goal of the ANN approach was to solve problems in the same way that a human brain would. However, over time, attention moved to performing specific tasks, leading to deviations from biology. Artificial neural networks have been used on a variety of tasks, including computer vision, speech recognition, machine translation, social network filtering, playing video games and medical diagnosis.[4]



IV. APPLICATION

Medical diagnosis plays a vital role to execute complicated tasks 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]

V. CONCLUSION

Recognition of the disease is mainly the purpose of the proposed technique which can recognize the heart diseases with little computational effort. The techniques can be used for the medical applications like detection and classification of diseases of heart with suitable classifier providing feasible approach for diagnosis. It also addresses how the disease analysis is possible for the heart disease detection which can be effectively detected in the early stage before it will led to disastrous consequences.

REFERENCES

[1] Prediction of heart disease using a hybrid technique in data mining classification Ankita Dewan ; Meghna Sharma 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)

[2] Heart Diseases Prediction with Data Mining and Neural Network Techniques Bandarage Shehani Sanketha Rathnayakc ; Gamage Upeksha Ganegoda 2018 3rd International Conference for Convergence in Technology (I2CT) Year: 2018

[3] Predictione Analysis of Classification Approaches for Heart Disease Prediction S. M. M. Hasan ; M. A. Mamun ; M. P. Uddin ; M. A.

Hossain 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2) Year: 2018

[4] Analysis of Neural Networks Based Heart Disease Prediction System SP Rajamhoana ; C. Akalya Devi ; K. Umamaheswari ; R. Kiruba ; K. Karunya ; R.Deepika 2018 11th International Conference on Human System Interaction (HSI) Year:2018

[5] Prediction of Cardiac Disease Based on Patient's Symptoms N. Prabakaran; R. Kannadasan 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) Year: 2018

[6] Heart disease prediction system based on hidden naive bayes classifier M. A.Jabbar ;Shirina Samreen 2016 International Conference on Circuits, Controls, Communications and Computing (I4C) Year: 2016