ABSTRACT

In the medical field, the diagnosis of heart disease is the most difficult task. The diagnosis of heart disease is difficult as a decision relied on grouping of large clinical and pathological data. Due to this complication, the interest increased in a significant amount between the researchers and clinical professionals about the efficient and accurate heart disease prediction. In case of heart disease, the correct diagnosis in early stage is important as time is the very important factor. Heart disease is the principal source of deaths widespread, and the prediction of heart disease is significant at an untimely phase. Machine learning in recent years has been the evolving, reliable and supporting tools in medical domain and has provided the greatest support for predicting disease with correct case of training and testing. The main idea behind this work is to study diverse prediction models for the heart disease and selecting important heart disease feature using Random Forests algorithm. Random Forests is the Supervised Machine Learning algorithm which has the high accuracy compared to other Supervised Machine Learning algorithms such as logistic regression etc. By using Random Forests algorithm, we are going to predict if a person has heart disease or not.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

The heart is a kind of muscular organ which pumps blood into the body and is the central part of the body's cardiovascular system which also contains lungs. Cardiovascular system also comprises a network of blood vessels, for example, veins, arteries, and capillaries. These blood vessels deliver blood all over the body. Abnormalities in normal blood flow from the heart cause several types of heart diseases which are commonly known as cardiovascular diseases (CVD). Heart diseases are the main reasons for death worldwide. According to the survey of the World Health Organization (WHO), 17.5 million total global deaths occur because of heart attacks and strokes. More than 75% of deaths from cardio-vascular diseases occur mostly in middle-income and low-income countries. Also, 80% of the deaths that occur due to CVDs are because of stroke and heart attack. Therefore, prediction of cardiac abnormalities at the early stage and tools for the prediction of heart diseases can save a lot of life and help doctors to design an effective treatment plan which ultimately reduces the mortality rate due to cardiovascular diseases.

Due to the development of advance healthcare systems, lots of patient data are nowadays available (i.e., Big Data in Electronic Health Record System) which can be used for designing predictive models for cardiovascular diseases. Data mining or machine learning is a discovery method for analyzing big data from an assorted perspective and encapsulating it into useful information. "Data Mining is a non-trivial extraction of implicit previously unknown and potentially useful information about data". Nowadays, a huge amount of data pertaining to disease diagnosis, patients etc. are generated by healthcare industries. Data mining provides a number of techniques which discover hidden patterns or similarities from data.

Therefore, in this paper, a machine learning algorithm is proposed for the implementation of a heart disease prediction system which was validated on two open access heart disease prediction datasets.

Data mining is the computer-based process of extracting useful information from enormous sets of databases. Data mining is most helpful in an explorative analysis because of nontrivial information from large volumes of evidence.
Medical data mining has great potential for exploring the cryptic patterns in the data sets of the clinical domain.

These patterns can be utilized for healthcare diagnosis. However, the available raw medical data are widely distributed, voluminous and heterogeneous in nature. This data needs to be collected in an organized form. This collected data can be then integrated to form a medical information system. Data mining provides a user-oriented approach to novel and hidden patterns in the Data The data mining tools are useful for answering business questions and techniques for predicting the various diseases in the healthcare field. Disease prediction plays a significant role in data mining. This paper analyzes the heart disease predictions using classification algorithms. These invisible patterns can be utilized for health diagnosis in healthcare data.

Data mining technology affords an efficient approach to the latest and indefinite patterns in the data. The information which is identified can be used by the healthcare administrators to get better services. Heart disease was the most crucial reason for victims in the countries like India, United States. In this project we are predicting the heart disease using classification algorithms. Machine learning techniques like Classification algorithms such as Random Forest, Logistic Regression are used to explore different kinds of heart-based problems.

CHAPTER 2

LITERATURE SURVEY

Machine Learning techniques are used to analyze and predict the medical data information resources. Diagnosis of heart disease is a significant and tedious task in medicine. The term heart disease encompasses the various diseases that affect the heart. The exposure of heart disease from various factors or symptom is an issue which is not complimentary from false presumptions often accompanied by unpredictable effects. The data classification is based on Supervised Machine Learning algorithm which results in better accuracy. Here we are using the Random Forest as the training algorithm to train the heart disease dataset and to predict the heart disease. The results showed that the medicinal prescription and designed prediction system is capable of prophesying the heart attack successfully. Machine Learning techniques are used to indicate the early mortality by analyzing the heart disease patients and their clinical records (Richards, G. et al., 2001). (Sung, S.F. et al., 2015) have brought about the two Machine Learning techniques, k-nearest neighbor model and existing multi linear regression to predict the stroke severity index (SSI) of the patients. Their study show that k-nearest neighbor performed better than Multi Linear Regression model. (Arslan, A. K. et al.,2016) have suggested various Machine Learning techniques such as support vector machine (SVM), penalized logistic regression (PLR) to predict the heart stroke. Their results show that SVM produced the best performance in prediction when compared to other models. Boshra Brahmi et al, [20] developed different Machine Learning techniques to evaluate the prediction and diagnosis of heart disease. The main objective is to evaluate the different classification techniques such as J48, Decision Tree, KNN and Naïve Bayes. After this, evaluating some performance in measures of accuracy, precision, sensitivity, specificity is evaluated.

Data source:

Clinical databases have collected a significant amount of information about patients and their medical conditions. Records set with medical attributes were obtained from the Cleveland Heart Disease database. With the help of the dataset, the patterns significant to the heart attack diagnosis are extracted.

The records were split equally into two datasets: training dataset and testing dataset. A total of 303 records with 76 medical attributes were obtained. All the attributes are numeric-valued. We are working on a reduced set of attributes, i.e., only 14 attributes.

All these restrictions were announced to shrink the digit of designs, these are as follows:

1. The features should seem on a single side of the rule.
2. The rule should distinct various features into the different groups.
3. The count of features available from the rule is organized by medica history people having heart disease only.

The following table shows the list of attributes on which we are working.

Table 2.1: List of Attributes

S no	Attribute Name	Description
1	Age	age in years
2	Sex	(1 = male; 0 = female)
3	Cp	Chest Pain
4	Trest bps	resting blood pressure (in mm Hg on admission to the hospital)
5	Chol	serum cholesterol in mg/d

6	Fbs	(Fasting blood sugar >120 mg/dl) (1 = true; 0 = false)
7	Restecg	Resting electrocardiographic results
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina (1=yes;0=no)
10	Old peak	ST depression induced by exercise relative to rest
11	Slope	The slope of the peak exercise ST segment
12	Ca	Number of major vessels (0-3) colored by fluoroscopy
13	Thal	3 = normal; 6 = Fixed defect; 7 = reversible fluoroscopy
14	Target	1 or 0

CHAPTER 3
AIM AND SCOPE OF PRESENT INVESTIGATION

3.1 EXISTING SYSTEM:

Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. There are many ways that a medical misdiagnosis can present itself. Whether a doctor is at fault, or hospital staff, a misdiagnosis of a serious illness can have very extreme and harmful effects. The National Patient Safety Foundation cites that 42% of medical patients feel they have had experienced a medical error or missed diagnosis. Patient safety is sometimes negligently given the back seat for other concerns, such as the cost of medical tests, drugs, and operations. Medical Misdiagnoses are a serious risk to our healthcare profession. If they continue, then people will fear going to the hospital for treatment. We can put an end to medical misdiagnosis by informing the public and filing claims and suits against the medical practitioners at fault.

Disadvantages:

- Prediction is not possible at early stages.
- In the Existing system, practical use of collected data is time consuming.
- Any faults occurred by the doctor or hospital staff in predicting would lead to fatal incidents.
- Highly expensive and laborious process needs to be performed before treating the patient to find out if he/she has any chances to get heart disease in future.

3.2 PROPOSED SYSTEM:

This section depicts the overview of the proposed system and illustrates all of the components, techniques and tools are used for developing the entire system. To develop an intelligent and user-friendly heart disease prediction system, an efficient software tool is needed in order to train huge datasets and compare multiple machine learning algorithms. After choosing the robust algorithm with best accuracy and performance measures, it will be implemented on the development of the smartphone-based application for detecting and predicting heart disease risk level. Hardware components like Arduino/Raspberry Pi, different biomedical sensors, display monitor, buzzer etc. are needed to build the continuous patient monitoring system.

3.3 FEASIBILITY STUDY:

A Feasibility Study is a preliminary study undertaken before the real work of a project starts to ascertain the likely hood of the project's success. It is an analysis of possible alternative solutions to a problem and a recommendation on the best alternative.

3.3.1 Economic Feasibility:

It is defined as the process of assessing the benefits and costs associated with the development of project. A proposed system, which is both operationally and technically feasible, must be a good investment for the organization. With the proposed system the users are greatly benefited as the users can be able to detect the fake news from the real news and are aware of most real and most fake news published in the recent years. This proposed system does not need any additional software and high system configuration. Hence the proposed system is economically feasible.

3.3.2 Technical Feasibility:

The technical feasibility infers whether the proposed system can be developed considering the technical issues like availability of the necessary technology, technical capacity, adequate response and extensibility. The project is decided to build using Python. Jupyter Notebook is designed for use in distributed environment