IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 08, Issue 9 (September. 2018), ||V (II) || PP 22-27

A Survey on Data Mining Techniques for Prediction of Heart Diseases

Susmitha K [1], B. Senthil Kumar[2]
M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India [1]
Assistant Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu,India$_2$
Corresponding Author: Susmitha K

Abstract: Heart disease (HD) is a disease of the heart or blood vessels, which causes death. In recent scenario, health issues are huge, due to this nature predicting and classifying into different conditions are very tedious. The field of data mining has involved in those domains to predict and to classify the abnormality along with its risk level. The previous studies have used several features to diagnosis the disease, which has been collected from patients. By applying different data mining algorithms, the patient data can be used for diagnosis as training samples. The main drawbacks of the previous studies are that need accurate and more number of features. This paper surveys about the recent data mining techniques applied for predicting heart diseases.

Index terms- Heart disease prediction, Data mining, Clustering, and Classification

---

Date of Submission: 31-08-2018 Date of acceptance: 15-09-2018

I. INTRODUCTION

Heart disease is the biggest cause of death nowadays. Blood pressure, cholesterol, pulse rate are the major reason for the heart disease. Some non-modifiable factors are also there. Such as smoking, drinking also reason for heart disease. The heart is an operating system of our human body. If the function of heart is not done properly means, it will affect other human body part also. Some risk factors of heart disease are Family history, High blood pressure, Cholesterol, Age, Poor diet, Smoking. When blood vessels are overstretched, the risk level of the blood vessels is increased. This leads to the blood pressure. Blood pressure is typically measured in terms of systolic and diastolic. Systolic indicates the pressure in the arteries when the heart muscle contracts and diastolic indicates the pressure in the arteries when the heart muscle is in resting state. The level of lipids or fats increased in the blood are causes the heart disease. The lipids are in the arteries hence the arteries become narrow and blood flow is also become slow. Age is the non-modifiable risk factor which also a reason for heart disease. Smoking is the reason for 40% of the death of heart diseases. Because it limits the oxygen level in the blood then it damage and tighten the blood vessels.

Various data mining techniques such as Naïve Bayes, KNN algorithm, Decision tree, Neural Network are used to predict the risk of heart disease [1]. The KNN algorithm uses the K user defined value to find the values of the factors of heart disease. Decision tree algorithm is used to provide the classified report for the heart disease. The Naïve Bayes method is used to predict the heart disease through probability. The Neural Network provides the minimized error of the prediction of heart disease. In all this above mentioned techniques the patient records are classified and predicted continuously. The patient activity is monitored continuously, if there is any changes occur, and then the risk level of disease is informed to the patient and doctor. The doctors are able to predict heart diseases at an earlier stage because of machine learning algorithms and with the help of computer technology.

Heart Disease: The heart is important organ or part of our body. Life is itself dependent on efficient working of heart. If operation of heart is not proper, it will affect the other body parts of human such as brain, kidney etc. It is nothing more than a pump, which pumps blood through the body. If circulation of blood in body is inefficient the organs like brain suffer and if heart stops working altogether, death occurs within minutes. Life is utterly reliant on proficient operation of the heart. The term Heart disease refers to illness of heart & blood vessel system in it. There are number of factors that amplifies the risk of Heart disease [2] such as Family history of heart disease, Smoking, Cholesterol, Poor diet, High blood pressure, High blood cholesterol, Obesity, Physical inactivity, Hyper tension.

Symptoms of a Heart Attack [3]:
Symptoms of a heart attack are Discomfort, pressure, heaviness, or pain in the chest, arm, or below the breastbone. Anxiety burning at back, jaw, throat, or arm. Fullness, indigestion, or choking feeling (may feel like heartburn). Some of the common indications are Sweating, nausea, vomiting, or dizziness that also includes extreme weakness, anxiety, or shortness of breath, rapid or irregular heartbeats.

Types of Heart diseases:
Sudhakar et al., Heart disease is wide term that comprises all types of diseases distressing various components of the heart. Heart denotes 'cardio'; therefore, entire heart diseases fit in to the category of cardiovascular diseases. Some types of Heart diseases are

1. **Coronary heart disease:**
   It also known as Heart disease (CAD), it is the most common type of heart disease across the world. It is a condition in which plaque deposits block the coronary blood vessels leading to a reduced supply of blood and oxygen to the heart.

2. **Arrhythmias:**
   It is connected with a disorder in the recurring movement of the heartbeat. The heartbeat can be slow, fast, or irregular. The unusual heartbeats are resulted by a short circuit in the heart's electrical system.

3. **Congestive heart failure:**
   It is a condition where the heart cannot pump enough blood to the rest of the body. It is commonly known as heart failure.

4. **Congenital heart disease :**
   It also known as congenital heart defect, it refers to the formation of an abnormal heart due to a defect in the structure of the heart or its functioning. It is also a type of congenital disease that children are born with.

5. **Cardiomyopathy:**
   It is the weakening of the heart muscle or a change in the structure of the muscle due to inadequate heart pumping. Recurrent of cardiomyopathy are hypertension, alcohol consumption, viral infections, and genetic defects.

6. **Angina pectoris:**
   It is a medical tenure for chest pain which is transpired due to inadequate contributes of blood to the heart, meanwhile termed as angina; it is a caution indicator for heart attack. The chest pain is at interval ranging for few seconds or minutes.

7. **Myocarditis:**
   It is an irritation of the heart muscle usually reasoned by viral, fungal, and bacterial contagion that influences the heart. It is an uncommon disease with few indications like joins pain, leg swelling or fever that cannot be directly related to the heart.

II. LITERATURE REVIEW:

This paper [1] authors had presented the feasibility study and the progress of heart disease classification embedded system. It provides a time diminution on electrocardiogram – ECG signal which can be practiced by decreasing the amount of data samples, without any significant loss. The objective of the urbanized system is the study of heart signals. The ECG signals are subjected onto the system that executes a preliminary filtering, and then utilizes a Gustafson–Kessel fuzzy clustering algorithm in order to exert for signal organization and correlation. The classification denotes usual heart diseases such as angina, myocardial infarction and coronary artery diseases. The system could also be used sudden "on duty" physicians, of any area of expertise, and could afford the first, or initial diagnose of any cardiopathy. If any system detects a heart problem, this system endows with better disease diagnose *PPV* evaluated to other testimonies, and therefore it tenders elevated assurance than other methods. Another foremost contemplation is the reality that this system was analogous to many other systems by accessing full data set, and this system exercised fuzzy clustering algorithm in order to diminish the data set, thus mitigating its use.

Data mining [2], as a resolution to haul out hidden pattern from the scientific dataset is projected to a database in this research. The database consists of 209 occurrences and 8 attributes. The system

was employed in WEKA and MATLAB software and prophecy accuracy within Apriori algorithm in just 3 steps, are compared. MATLAB is pioneer as better performance software. Wide ranges of Apriori algorithms' sturdy system in data mining were evaluated to predict heart disease. A sole model consisting of one filter and appraisal methods are evolved. Three strong rules, as well as different estimation methods, are applied to find the superior software. Apriori rules are measured concerning their actual number of support, better accuracy, and considering strong rules. The high-performance software was introduced. The experiment can serve as a realistic tool for physicians to in effect predict uncertain cases and recommends consequently.

Authors in [3] presented a proficient advance for the forecast of heart attack from the heart disease database. Initially, the heart disease database is huddled using the K-means clustering algorithm, which will extort the data appropriate to heart attack from the database. This approach permits expertise the number of fragments through its k

parameter. Consequently the frequent patterns are excavated from the extracted data, relevant to heart disease, using the MAFIA (Maximal Frequent Item set Algorithm) algorithm. The machine learning algorithm is modeled with the selected major patterns for the effectual prediction of heart attack. They have engaged the ID3 algorithm as the training algorithm to prove level of heart attack with the decision tree. The results showed that the designed prediction system is competent of forecasting the heart attack effectively.

In this paper [4] authors described about a prototype using data mining techniques mainly Naïve Bayes and WAC (Weighted Associated Classifier). The dataset is composed of important factors such as age, sex, diabetic, height, weight, blood pressure, cholesterol, fasting blood sugar, hypertension, disease. The system indicates whether patient had a risk of heart disease or not.

In this paper [5], authors proposed confidential scheme for predicting heart disease using two different models, Naive Bayes and Logistic Regression. As identified through survey, it is a need to have combinational approach to increase the accuracy of prediction for heart disease.

In this paper [6] authors proposed that, heart disease is one of the major causes of demise in the region of the world and it is essential to forecast the disease at a precipitate phase. The computer aided systems assists the doctor as a gizmo for forecasting and establishing heart disease. The intention of this paper is to extend about Heart related cardiovascular disease and to brief about accessible decision support systems for the computation and study of heart disease continued by data mining and hybrid intelligent techniques. Many DSS remains to predict the heart disease with several methodologies. The World life expectation statistics involve that heart disease has extended more in number. So it is essential to construct an efficient intelligent trusted automated system which predicts the heart disease precisely based on the symptoms according to gender/age and province knowledge of experts in the field at the lowest cost.

Authors in this paper [7] explicate that figures reveal that a heart disease is one of the foremost factors behind deaths throughout the world. Data mining techniques are pretty effectual in manipulative scientific support systems and having the capability to determine hidden patterns and relationships in medical data. Till now, Data mining classification techniques is applied to examine the various kinds of heart based problems. This paper is intended at mounting a heart disease prediction system using data mining clustering methods. This paper crews the various clustering techniques, k-mean, EM and the farthest first algorithm for the prophecy of heart disease. End result proves that farthest first clustering algorithm is the finest algorithm as evaluate to other algorithms. Since the ratio of correctly classified occurrences to the cluster is highest and the time taken to construct the model is minimum. This system can be further extended. More number of input attributes can be used and it can be further expanded by escalating the no. of the clusters. The same experiment can also be performed on other data mining tool such as R. And also the ensemble of classifiers can also be done to estimate their performance with the unique classifiers. Above algorithms can be subjected to other datasets in order to scrutinize whether the identical algorithm gives the highest precision or not.

Authors in this paper [8] proposed the incorporation of accessing a clustering approach and regression methodology. The clustering approach used is DBSCAN and for regression, multiclass logistic regression is subjected. By executing DBSCAN clustering algorithm, the entire dataset is fragmented into disjoint clusters. Resulted clusters were found to enclose fewer occurrences are then taken for consideration .These clusters are focused to multiclass logistic regression. This result is due to the clustering approach acquired by an unsupervised process. Once regression is achieved, we have accomplished at a termination, about actual variety of cardiac arrhythmia it is. The projected method accomplishes an overall accuracy of 80%, when evaluated with various other existing approaches. It projects a method for the prophecy of type of cardiac arrhythmia by assembling the use of DBSCAN clustering and multi class logistic regression algorithms. By balancing PCA-CRA with other methods, this method is found to be 80% accurate.

This paper [9] intends that large data existing from medical diagnosis is scrutinized by means of data mining tools and valuable information known as knowledge is hauling out. Mining is a method of investigating colossal sets of data to acquire the patterns which are hidden and formerly unknown associations and knowledge

detection to facilitate the enhanced understanding of medical data to thwart heart disease. There are several DM techniques available namely Classification techniques concerning Naïve bayes (NB), Decision tree (DT), Neural network (NN), Genetic algorithm (GA), Artificial intelligence (AI) and Clustering algorithms like KNN, and Support vector machine (SVM). Numerous studies have been conceded out for mounting prophecy model by accessing entity technique and also by coalescing two or more techniques. This paper offers a rapid and simple evaluation and perceptive of obtainable prophecy models by means of data mining from 2004 to 2016. In this paper, a survey conducted from 2004 to 2015 gives the scheme of various models obtainable and the various data mining methodologies used. The exactness gained with these models is also specified. It is pragmatic that all the techniques accessible have not use big data analytics. Exploiting of big data analytics along with data mining will offer talented results to get the finest precision in manipulating the prophecy model.

Authors in this paper [10] recommends that heart disease is one of the diseases due to that fatality will occur mostly, and according to the world health organization the percentage is high for that. So Heart disease is determined for the big Data approach, and as Big Data is measured so use Hadoop Map diminish platform. For clustering improved K-Means and for the classification principle decision tree algorithm i.e. ID3 can be accessed in the hybrid approach. As second estimation is too better, the system is very helpful for the facilitating the forecast methods, based on the some restrictions like chest pain, cholesterol, age, resting Bp, Thalac and many more. Due to this system medical decision making will be enhanced as well as being rapid. It's also will impact on the humanizing the treatment process. In such way it will be very helpful in the prophecy of the heart disease. In such way authors had cultured about the big data and its properties, with its disputes and concerns. In the medical field the various parameters individuals are affecting to the heart. Improved K-Means is the algorithm which is viewing the precision in the centroid assortment more than the simple K-Means.

Authors in paper [11] intended that the medical doings examination plays a significant role in present trend. Discovery and study of medical doings is the most vital concern in real time scenario, since the requirement of training samples and adequate data's formulate these procedures much complex. This medical data analysis can be executed by effectual data mining methodology and advances. There are numerous unlike methods to diagnosis and prognosis diabetes mellitus. This paper had presented diverse techniques of the data mining methodologies to resolve the diabetes disease diagnosis problem. From the analysis, the discovery of several problems has been mentioned and locates in clinical datasets handling process.

In paper [12], a survey on a range of methodologies and algorithms for efficient classifier in premises of two issues such as class imbalance and dimensionality reduction has implemented. In the research it is noticed that numerous work categorizes under the class imbalance problem reduction and dimensionality reduction problems, but there is not a bit of the accesses were determined on both issues. So creating a classifier for the high-dimensional data with class imbalance problem will be a fascinating region for the future research. Numerous class imbalance problems are still not adequate for multiple class imbalance problems. This survey presents an outline on dimensionality reduction and class imbalance classification with the probable issues and outcomes. It portrays the major issues that slow down the classifier conduct due to these two problems.

Table 1.0: Comparison table

Paper Number	Technique	Advantages	Disadvantages
1	Filtering process and fuzzy cluster algorithm	Validates the principles of embedded system, and promotes the maximum possible efficiency	Hardware limitations i.e., lack of memory management, absence of operational system
2	Apriori algorithm applying WEKA and MATLAB software	High performance which offers better accuracies	Very slow and desires candidate generation every time
3	K-means clustering algorithm with MAFIA method	Easy to implement with a large number of variables, K-Means may be computationally faster than hierarchical clustering.	Difficult to predict the number of clusters (K-Value), The order of the data has an impact on the final results.
4	Naïve Bayes and WAC (Weighted Associated Classifier)	Very simple and easy to use, highly scalable, make probabilistic predictions	Features in the output class are independent, scarcity of data
5	ɔgistic Regression and /M	Solve privacy violation problem with high accuracy	Takes more speed and time for training and testing, discrete data will lead to some other problems

6	Data mining algorithms such as Neural networks, naïve bayes, decision tree, genetic algorithm	Effectual way to tackle the risks, integrates the strengths of various techniques	Only 80% of accuracy can be achieved, lack of certain data security
7	Clustering techniques such as k-mean, EM and the farthest first algorithm	Automatic recovery from failure	Complexity and inability to recover from database corruption.
8	Clustering approach with DBSCAN methodologies	Different link connection is minimized, robust	Not entirely deterministic, border points that are reachable from more than one cluster can be part of either cluster; quality depends upon the distance of the function region query.
9	Classification techniques involving Naïve bayes (NB), Decision tree (DT), Neural network (NN), Genetic algorithm (GA), Artificial intelligence (AI) and Clustering algorithms like KNN, and Support vector machine (SVM).	User friendly and scalable with 90% of accuracy	All the techniques available have not used big data analytics
10	Improved K-means and ID3	Comparison of time complexity and accuracies is possible, speed and ease of use.	Sensitive to the selection of initial cluster center, usually end without global optimal solution, but suboptimal solution; Sometimes the result of cluster may lose balance
11	Support Vector Machine (SVM)	Maximizes the prediction accuracy, avoids over-fitting problem, classifies and diagnosis effectively	Needs more tuning parameters and deep study is necessary, Can't perform using statistical analysis, Group attention selection process needs more attention.
12	Principal Component Analysis (PCA)	Low noise sensitivity, decreased requirements for capacity and memory, and increased efficiency	Only be used if the original variables are correlated and homogeneous

The above table 2.0. Depicts the working methodologies of various data mining techniques which can be used to achieve prediction of heart diseases.s

III. CONCLUSION

Heart disease is one of the major problems in nowadays which leads to causality. Predicting heart diseases is possible only by the consideration of attributes. This analyzing method of the attributes can be achieved by the inclusion of data mining techniques. Data mining methodologies embraces methods such as Neural networks, naïve bayes, clustering mechanisms, classification, big data, etc. Further implementation has to be done in order to predict heart disease in a big data environment.

REFERENCES

[1]. de Carvalho Junior, Helton Hugo, et al. "A heart disease recognition embedded system with fuzzy cluster algorithm." *Computer methods and programs in biomedicine* 110.3 (2013): 447-454.

[2]. Mirmozaffari, Mirpouya, Alireza Alinezhad, and Azadeh Gilanpour. "Data Mining Apriori Algorithm for Heart Disease Prediction." Int'l Journal of Computing, Communications & Instrumentation Engg (IJCCIE) 4.1 (2017).

[3]. Khaing, Hnin Wint. "Data mining based fragmentation and prediction of medical data." *Computer Research and Development (ICCRD), 2011 3rd International Conference on*. Vol. 2. IEEE, 2011.

[4]. Patel, Ajad, Sonali Gandhi, Swetha Shetty, and Bhanu Tekwani. "Heart Disease Prediction Using Data Mining." (2017).

[5]. Wghmode, Mr Amol A., Mr Darpan Sawant, and Deven D. Ketkar. "Heart Disease Prediction Using Data mining Techniques." *Heart Disease* (2017).

[6]. Vijayashree, J., and N. Ch SrimanNarayanaIyengar. "Heart disease prediction system using data mining and hybrid intelligent techniques: A review." Int. J. Bio-Sci. Biotechnol 8 (2016): 139-148.

[7]. Singla, Meenu, and Kawaljeet Singh. "Heart Disease Prediction System using Data Mining Clustering Techniques."

[8]. Cp, Prathibhamol, Anjana Suresh, and Gopika Suresh. "Prediction of cardiac arrhythmia type using clustering and regression approach (P-CA-CRA)." Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on. IEEE, 2017.

[9]. Banu, NK Salma, and Suma Swamy. "Prediction of heart disease at early stage using data mining and big data analytics: A survey." Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 016 International Conference on. IEEE, 2016.

[10]. Mane, Tejaswini U. "Smart heart disease prediction system using Improved K-means and ID3 on big data." Data Management, Analytics and Innovation (ICDMAI), 2017 International Conference on. IEEE, 2017.

[11]. Senthil Kumar, B., and Dr Gunavathi R. "A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis." *IJARCCE* 5 (2016): 463-467.

[12]. Kumar, B. Senthil, and R. Gunavathi. "Comparative and Analysis of Classification Problems." *Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org* 7.8 (2017).

Susmitha K "A Survey on Data Mining Techniques for Prediction of Heart Diseases. " IOSR Journal of Engineering (IOSRJEN), vol. 08, no. 9, 2018, pp. 22-27.