IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

A Survey on Heart Disease Prediction Using Data Mining Techniques

G. Srinaganya[1*], A. Kiruba[2]

[1,2] Department of Computer Science, Shrimathi Indira Gandhi college, Tiruchirappalli-2

[*]*Corresponding Author: srinaganyag1978@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i5.877880 | Available online at: www.ijcseonline.org

Accepted: 15/May/2019, Published: 31/May/2019

Abstract— The health care environment is found to be rich in information, but poor in extracting knowledge from the information. This is because of the lack of effective analysis tool to discover hidden relationships and trends in them. By applying the data mining techniques, valuable knowledge can be extracted from the health care system. Heart disease is a group of condition affecting the structure and functions of the heart and has many root causes. Heart disease is the leading cause of death in the world over past ten years. Researches have been made with many hybrid techniques for diagnosing heart disease. This paper deals with an overall review of the application of data mining in heart disease prediction.

Keywords— Cardio Vascular Disease, Data Mining, Feature Selection, Classification, Association Rule Mining, Clustering

I. INTRODUCTION

The heart is an important organ of all living creatures, which plays a vital role of pumping blood to the rest of the organs through the blood vessels of the circulatory system. Any functional problem in the heart has a direct impact on the survival of concern human being, since it affects other parts of the body such as brain, lungs, kidney, liver etc. Heart Diseases describe a range of conditions that affect the heart and stand as a leading cause of death all over the world. The clinical symptoms of the Heart Disease complicate the prognosis, as it is influenced by many factors like functional and pathologic appearance. This could subsequently delay the prognosis of the disease. Hence, there is a need for the invention of newer concepts to improve the prediction accuracy with short span. Disease prognosis through numerous factors or symptoms is a multi-layer problem, even that could lead to a false assumption. Therefore, an attempt is made to bridge the knowledge and the experience of the experts and to build a system that fairly supports the diagnosing process.

Data Mining is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistics [1]. Data mining techniques are the result of a long process of research and product development. Data Mining is divided into two tasks, such as Predictive Tasks and Descriptive Tasks. Productive Tasks predict the value of a specific attribute based on another attribute. Classification, Regression and Deviation Deduction come under Predictive Tasks. Descriptive Tasks derive pattern that summarizes the relationship between data. Clustering, Association Rule Mining and Sequential Pattern Discovery are coming under Descriptive Tasks. Data Mining involves few steps from raw data collection to some form of new knowledge. The iterative process consists of the following steps like Data cleaning, Data Integration, Data Selection, Data transformation, Data Mining, Pattern Evaluation, and Knowledge Representation.

Medical Data Mining is a domain of challenge which involves a lot of imprecision and uncertainty. Provision of quality services at affordable cost is the major challenge faced in the health care organization. The poor clinical decision may lead to disastrous consequences. Health care data is massive. Clinical decisions are often made based on the doctor's experience rather than on the knowledge rich data hidden in the database. This in some cases will result in errors, excessive medical cost which affects the quality of service to the patients [2]. Medical history data comprise of a number of tests essentials to diagnose a particular disease. It is possible to gain the advantage of Data mining in health care by employing it as an intelligent diagnostic tool. The researchers in the medical field identify and predict the disease with the aid of Data mining techniques [21]. In this paper, the overview of the heart disease is depicted in the section 2, section 3 presents the Data Mining techniques, section 4 detailed the literature review on the Prediction of Heart Disease using Data Mining techniques. Section 5 depicts the overall summary of this paper.

II. HEART DISEASE

The initial diagnosis of a heart attack is made by a combination of clinical symptoms and characteristic electrocardiogram (ECG) changes. An ECG is a recording of the electrical activity of the heart. Confirmation of a heart attack can only be made hours later through detection of elevated creatinine phosphokinase (CPK) in the blood. CPK is a muscle protein enzyme which is released into the blood circulation by dying heart muscles when their surrounding dissolves [3]. World Health Organization in the year 2003 reported that 29.2% of total global deaths are due to Cardio Vascular Disease (CVD). By the end of this year, CVD is expected to be the leading cause for deaths in developing countries due to change in life style, work culture and food habits. Hence, more careful and efficient methods of cardiac diseases and periodic examination are of high importance [1]

III. DATA MINING TECHNIQUES

Data Mining is the process of extracting valid, authentic, and actionable information from large databases. Data Mining is an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications. Data mining strategies fall into two broad categories namely Supervised Learning and Unsupervised Learning. Supervised Learning methods are deployed when there exists a field or variable (target) with known values and about which predictions will be made by using the values of other fields or variables (inputs). Unsupervised Learning methods tend to be deployed on data for which there do not exist a field or variable with known values, while fields or variables do exist for other fields or variables.

FEATURE SELECTION

Feature selection is a process used in machine learning in which a subset of the features available from the data is selected for application of a learning algorithm. It is necessary because it is computationally not feasible to use all available features or because of problems of estimation when limited data samples are present. Feature selection from the available data is vital to the effectiveness of the methods employed. Extracted features can be ranked with respect to their contribution and utilized accordingly. Existing feature selection methods for machine learning typically fall into two broad categories; those which evaluate the worth of features using the learning algorithm that is to be ultimately applied to the data, and those which evaluate the worth of features by using heuristics based on general characteristics of the data. The former is referred to as wrappers and the latter filters.

CLASSIFICATION TECHNIQUES

The classification task in machine learning is to take each instance of a dataset and assign it to a particular class. A classification-based system attempts to classify all the patient either having heart disease or not. The challenge in this is to minimize the number of false positives and false negatives. Classification maps a data item into one of several predefined categories. These algorithms normally output "classifiers", for example, in the form of decision trees or rules. An ideal application in intrusion detection will be to gather sufficient "normal" and "abnormal" audit data for a user or a program, then apply a classification algorithm to learn a classifier that will determine (future) audit data as belonging to the normal class or the abnormal class. Five general categories of techniques have been tried to perform classification for intrusion detection purposes. They are inductive rule generation, GA, FL, Neural Networks (NN) and immunological based techniques.

CLUSTERING TECHNIQUES

Data clustering is a common technique for statistical data analysis which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. It is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets or clusters, so that the data in each subset share some common trait which is often proximity according to some defined distance measure.

Machine learning typically regards data clustering as a form of unsupervised learning. Clustering is useful in intrusion detection as malicious activity should cluster together, separating itself from non-malicious activity. Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters. Clustering provides some significant advantages over the classification techniques already discussed, in that it does not require the use of a labeled data set for training.

ASSOCIATION RULE MINING

Association rules are if/then statements that help to uncover relationships between unrelated data in a database, relational database or other information repository.

Association rules are used to find the relationships between the objects which are frequently used together. Applications of association rules are basket data analysis, classification, cross-marketing, clustering, catalog design, and loss-leader analysis etc.

IV. LITERATUTE REVIEW

Nahar, Jesmin, et al [1], investigates the sick and healthy factors which contribute to heart disease for males and females. Association rule mining, a computational intelligence approach, is used to identify these factors and the UCI Cleveland dataset, a biological database, is considered along with the three rule generation algorithms – Apriori, Predictive Apriori and Tertius.

Vijiyarani, S., and S. Sudha [2] this paper analyzes the classification tree techniques in data mining. The aim of this paper is to investigate the experimental results of the performance of different classification techniques for a heart disease dataset. The classification tree algorithms used and tested in this work are Decision Stump, Random Forest, and LMT Tree algorithm.

Gayathri, P., and N. Jaisankar [3] the survey of the papers related to heart disease and also the survey of many categories of heart disease such as coronary heart disease, coronary artery disease, heart failure, ischemic heart disease, cardiovascular disease, congenital heart disease, valvular heart disease and hypoplastic left heart syndrome are presented in this paper.

Shouman, Mai, Tim Turner, and Rob Stocker [4] the research presented here is part of work to develop tools to assist healthcare practitioners to diagnosis heart disease earlier in the hope of earlier interventions in this preventable killer. The relative accuracy of common data mining techniques in heart disease diagnosis is difficult to assess from the literature. This research investigates Decision Tree, Naïve Bayes, and K-nearest Neighbour performance in the diagnosis of heart disease patients.

Amato, Filippo, et al [5] this paper ANNs represent a powerful tool to help physicians perform diagnosis and other enforcements. In this regard, ANNs have several advantages including: (i) The ability to process large amount of data (ii) Reduced likelihood of overlooking relevant Information (iii) Reduction of diagnosis time.

Persi Pamela, I., and P. Gayathri [6] A fuzzy system is one of the soft computing methodologies is proposed in this paper along with a data mining technique for efficient diagnosis of coronary heart disease. Though the database has 76 attributes, only 14 attributes are found to be efficient for CHD diagnosis as per all the published experiments and doctors' opinion. So only the essential attributes are taken from the heart disease database. From these attributes crisp rules are obtained by employing CART decision tree algorithm, which are then applied to the fuzzy system. A Particle Swarm Optimization (PSO) technique is applied for the optimization of the fuzzy membership functions where the parameters of the membership functions are altered to new positions.

Chaurasia, Vikas, and Saurabh Pal [7] This research paper intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques which will be useful for medical practitioners to take effective decision. The objective of this research work is to predict more accurately the presence of heart disease with reduced number of attributes.

Thenmozhi, K., and P. Deepika [8] the authors explored the various decision tree algorithms like ID3, C4.5, C5.0 and J48 in the classification and prediction of heart disease.

Kim, Jae-Kwon, et al [9] This paper proposes the Fuzzy Rule-based Adaptive Coronary Heart Disease Prediction Support Model (FbACHD_PSM), which gives content recommendation to coronary heart disease patients. The proposed model uses a mining technique validated by medical experts to provide recommendations.

Seera, Manjeevan, and Chee Peng Lim [10] in this paper, a hybrid intelligent system that consists of the Fuzzy Min-Max neural network, the Classification and Regression Tree, and the Random Forest model is proposed, and its efficacy as a decision support tool for medical data classification is examined. The hybrid intelligent system aims to exploit the advantages of the constituent models and, at the same time, alleviate their limitations.

Bashir, Saba, Usman Qamar, and M. Younus Javed [11] The objective of the proposed research is to predict the heart disease in a patient more accurately. The proposed framework uses majority vote based novel classifier ensemble to combine different data mining classifiers. UCI heart disease dataset is used for results and evaluation.

Shabana, ASMI P., and S. Justin Samuel [12] Different data mining techniques such as Naive Bayes, Decision Tree, Linear Regression and Association Rule are used to predict the heart disease. Data mining techniques in all disease diagnosis applied over all disease treatment dataset investigate if hybrid data mining techniques can achieve equivalent (or better) results in identifying suitable treatments as that achieved in the diagnosis. In this paper, the proposed work is to more accurately predict the presence of heart disease with added attributes of the disease and using association rules.

Aljaaf, A. J., et al [13] In this study, a multi-level risk assessment of developing heart failure has been proposed, in which a five risk levels of heart failure can be predicted using C4.5 decision tree classifier. On the other hand, we are boosting the early prediction of heart failure through involving three main risk factors with the heart failure data set.

Bashir, Saba, Usman Qamar, and Farhan Hassan Khan [14] this research paper presents a novel classifier ensemble framework based on enhanced bagging approach with multi-objective weighted voting scheme for prediction and analysis

of heart disease. The proposed model overcomes the limitations of conventional performance by utilizing an ensemble of five heterogeneous classifiers: Naive Bayes, linear regression, quadratic discriminant analysis, instance-based learner and support vector machines.

Kim, Jaekwon, Jongsik Lee, and Youngho Lee [15] developed model for CHD prediction must be designed according to rule-based guidelines. In this study, a fuzzy logic and decision tree (classification and regression tree [CART])-driven CHD prediction model was developed for Koreans. Datasets derived from the Korean National Health and Nutrition Examination Survey VI (KNHANES-VI) were utilized to generate the proposed model.

V. CONCLUSION

The most important and difficult task in medicine is Medical diagnosis. The problem here is detecting a disease from several factors or symptoms, since it may lead to false assumptions with unpredictable results. Heart disease prediction is a major challenge in the healthcare industry. Instead of going for a number of tests, predicting heart disease with a smaller number of attributes is challenging task in Data Mining. Existing literature shows that Classification task in Data Mining plays a vital role in heart disease prediction when compared with Clustering, Association Rule and Regression.

REFERENCES

[1] Nahar, Jasmine, et al, "*Association rule mining to detect factors which contribute to heart disease in males and females*", Expert Systems with Applications, Vol. 40 Issue. 4, pp. 1086-1093, 2013.

[2] Vijiyarani, S., and S. Sudha, "*An efficient classification tree technique for heart disease prediction*", International Conference on Research Trends in Computer Technologies (ICRTCT-2013) Proceedings published in International Journal of Computer Applications (IJCA)(0975–8887). Vol. 201, 2013.

[3] Gayathri, P., and N. Jaisankar, "*Comprehensive study of heart disease diagnosis using data mining and soft computing techniques*", 2013.

[4] Shouman, Mai, Tim Turner, and Rob Stocker, "*Integrating clustering with different data mining techniques in the diagnosis of heart disease*", J. Comput. Sci. Eng, Vol. 20 Issue.1, 2013.

[5] Amato, Filippo, et al, "*Artificial neural networks in medical diagnosis*", pp. 47-58, 2013.

[6] Persi Pamela, I., and P. Gayathri, "*A fuzzy optimization technique for the prediction of coronary heart disease using decision tree*", 2013.

[7] Chaurasia, Vikas, and Saurabh Pal, "*Data mining approach to detect heart diseases*", 2014.

[8] Thenmozhi, K., and P. Deepika, "*Heart disease prediction using classification with different decision tree techniques*", International Journal of Engineering Research and General Science, Vol. 2, Issue. 6, pp. 6-11, 2014.

[9] Kim, Jae-Kwon, et al, "*Adaptive mining prediction model for content recommendation to coronary heart disease patients*", Cluster computing, Vol. 17, Issue. 3, pp. 881-891, 2014.

[10] Seera, Manjeevan, and Chee Peng Lim, "*A hybrid intelligent system for medical data classification*", Expert Systems with Applications, Vol. 41, Issue. 5, pp. 2239-2249, 2014.

[11] Bashir, Saba, Usman Qamar, and M. Younus Javed, "*An ensemble based decision support framework for intelligent heart disease diagnosis*", Information Society (i-Society), *2014 International Conference on*. IEEE, 2014.

[12] Shabana, ASMI P., and S. Justin Samuel, "*An analysis and accuracy prediction of heart disease with association rule and other data mining techniques*", Journal of Theoretical and Applied Information Technology, Vol. 79, Issue. 2, pp. 254-60, 2015.

[13] Aljaaf, A. J., et al, "*Predicting the likelihood of heart failure with a multi level risk assessment using decision tree*", Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), *2015 Third International Conference on*. IEEE, 2015.

[14] Bashir, Saba, Usman Qamar, and Farhan Hassan Khan, "*BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting*", Australasian physical & engineering sciences in medicine, Vol. 38, Issue. 2, pp. 305-323, 2015.

[15] Kim, Jaekwon, Jongsik Lee, and Youngho Lee, "*Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree*", Healthcare informatics research, Volume. 21, Issue. 3, pp. 167-174, 2015.

