A Review on Healthcare Sector using Data Mining and Classification Techniques

Sharli Das[1], Prof. Goldy Saini[2], Dr. Kalpana Rai[3], Dr. Sneha Soni[4]

[1]M. Tech Research Scholar, [2]Assistant Professor, [3,4]Professor
[1,2,3,4] Department of Computer Science & Engineering
[1,2,3,4] SIRT-E, Bhopal (M.P.), India

Abstract: The huge amount of medical data generated by healthcare devices are large and complex to be analyzed by traditional methods. Data mining is used to improve the process by discovering patterns and features is large complex data. Several techniques have been presented to give accurate medical diagnose for various diseases. In this work, we present a literature work for heart disease prediction using data mining, machine learning, and classification techniques.

Keywords: Data mining, Machine learning, Classification techniques, Health care, Prediction Systems.

Introduction

The amount of data in the medical industry is increasing day by day. It is a challenging task to handle a large amount of data and extracting productive information for effective decision making. For this reason, medical industry demands to apply a special technique which will provide fruitful decision from a vast database. Data mining is an exciting field of machine learning and thus capable of solving this type of problem very well. For solving various kinds of real-world problems, data mining is a novel field for discovering hidden patterns and the valuable knowledge from a large dataset. Because it is very strenuous to extract any useful information without mining large database. In brief, it is an essential procedure for analyzing data from various perspectives and gathering knowledge. However, heath care industry is another field where a substantial amount of data collected using different clinical reports and patients manifestations [5]. Nowadays, people can face any heart failure symptoms at any stage of a lifetime. But old people face this type of problem rather than the young people. Data mining classification techniques can discover the hidden relationship along correlated features which plays a consequential role in predicting the class label from a large dataset. By using those hidden patterns along with the correlated features, it is straightforward to detect heart disease patients without any support of medical practitioners. Then, it will act as an expert system for separating patients with heart disease and patients with no heart disease more accurately with lower cost and less diagnosis time.

Data mining plays an immense role in extracting useful information from big data. It is widely used in almost every field of life like medicine, engineering, business, and education. Data mining is used to explore the data to extract the hidden crucial decision making information from the collection of the past repository for future. A variety of machine learning algorithms have been used to understand the complexity and non-linear interaction between different factors by decreasing the error in prediction and factual outcomes. Due to ever increasing medical data, we need to leverage on machine learning algorithms to assist medical healthcare professionals in analyzing data and making accurate and precise diagnostic decisions. In medical data mining, different classification algorithms are used to predict the CVD in patients and death predictions due to the heart attack [13].

Machine learning models learn from patterns in given training examples without explicit instructions and then use inference to develop useful predictions [1, 2]. Classification methods are widespread in the medical area for identifying and predicting diseases more accurately. Diseases and health problems like liver cancer, chronic kidney, breast cancer, diabetes, and heart syndrome have a significant impact on one's health and may lead to death if ignored. By extracting hidden patterns and relationships in the database, the healthcare industry will make successful decisions. Due to progress in machine learning and artificial intelligence, several classifiers and clustering algorithms like K-nearest, Decision Tree, Random Forest, Support Vector Machine (SVM), Naïve Bayes, and other algorithms can give a solution to this situation.

In medical applications, the most famous machine learning technique is classification because it corresponds to problems appearing in everyday life. Classification algorithms first use the training data to build a model and then the resulting model is applied to the test data to obtain a prediction. Different classification methods have been implemented for disease diagnosis and the findings be very promising. These techniques can minimize diagnostic faults and results can be obtained in a short period of time [7].

II. Classification Techniques

Classification techniques are widely used in healthcare, since they are capable of processing large set of data. The common used techniques in healthcare are Naïve Bayesian, support vector machine, Nearest neighbor, decision tree, Fuzzy logic, Fuzzy based neural network, Artificial neural network, and genetic algorithms. Machine learning with classification can be efficiently applied in medical applications for complex measurements. Modern classification techniques provide more intelligent and effective prediction techniques for heart disease [4].

A. Machine Learning

Machine Learning is a technique that allows the computers to learn from the past data without being explicitly programmed. So, basically it is a discipline to predict things accurately using statistical methods and algorithms. Machine learning enables computers to manage new situations by analyzing, self-training, observations and experiences. Machine learning has found its application in numerous fields such as healthcare, agriculture, banking, natural language processing, speech recognition, optimisation, etc.

B. Deep Learning

Deep learning is a part of machine learning that helps to teach computers to do what comes naturally to humans: learn by examples and observations. Deep learning is a type of machine learning which is inspired by the structure and functioning of a human brain. Deep learning came up into the picture when the processing power of modern computers grew exponentially. Applications of deep learning involve automatic speech recognition, fraud detection, healthcare, etc.

C. Artificial Neural Networks (ANN)

It is an effort to simulate the network of neurons that make up a human brain in order to enable computers to learn things and take decisions just like humans. It consists of many layers and each layer has many neurons. Initially, an ANN undergoes a training phase where it learns to observe and recognize patterns in dataset, whether visually, aurally or textually. During this training phase, the network makes a comparison of its predicted output with actual output. The variation between both outcomes is adjusted using back propagation.

D. Data Mining

In general, data mining means selection and extraction of appropriate and useful information from vast amount of available data. The outcome and result of data mining is the knowledge and patterns extracted from the given data. There are three major steps involved in data mining: Data Pre- Processing, Data Extraction and Data Presentation. It can be applied in various fields like education, healthcare, banking, e-commerce, scientific analysis, etc.

E. K-Nearest Neighbor

K-Nearest Neighbor (KNN), a supervised learning model as well, is used to classify the test data using the training samples directly. It. is a method for classifying objects based on closest training data in the feature space. Otherwise, the class of a new sample is predicted based on some distance metrics where the distance metric can be a simple Euclidean distance. In the practical steps, KNN first calculates k (Number of the nearest neighbors), and it finds the distance between the training data and then sorts the distance. Subsequently, a class label will be assigned to the test data based on the majority voting.

F. Random Forest

Random Forest (RF) refers to ensembles of simple tree predictors, and each tree produce and outcome. For the regression model, the tree outcome is an estimate of the dependent value given the predictors. On the other hand, for the classification model, the tree outcome takes the form of a class membership that classifies a set of independent values given the predictors with one of the categories present in the dependent value.

G. Support Vector Machine

Support Vector Machine (SVM) is one of the standard set of supervised learning model employed in classification. It is defined as the finite-dimensional vector spaces where each dimension characterizes a feature of a particular sample. A support vector machine aims to find the best highest-margin separating hyperplane between the two classes. Due to its computational competence on big data sets SVM has been proved as an effective method in high-dimensional space problems.

H. Decision Tree

Decision Tree (DT), currently, is one of the most potent and popular classifications and prediction algorithms used in machine learning. It has been widely used to examine data and make decisions by many researchers as classifiers in the healthcare domain. DT creates a model that predicts the value of a target variable by learning simple decision rules inferred from the data features and splitting data into branch-like segments. The input values can be continuous or discrete. The leaf nodes return class labels or probability scores. In theory, the tree can be converted into decision rules. These rules of classification can easily be presented visually [2].

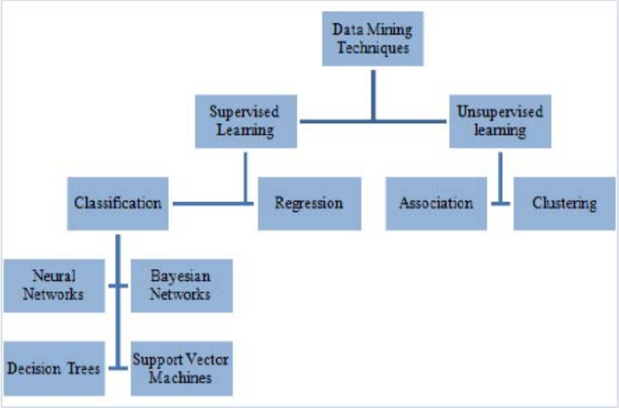

Figure 1: Classification of Data Mining Algorithms [6].

III. Literature Review

Different researchers had proposed different decision support systems to predict heart disease making use of various machine learning algorithms. Different decision support systems proposed by different researchers in predicting heart disease are discussed in this section.

In the last few decades, medical science has used the technological advancements very well to improve the quality of healthcare. These advancements in technology have paved ways for accurate diagnosis and prediction of diseases [1]. Various researchers have proposed a number of models for predicting heart diseases using different technologies such as artificial neural networks, machine learning, data mining, etc. This paper analyses the work done by various researchers on the accuracy of heart disease prediction through the different approaches. A detail literature review has been provided in the study. The analysis has also been presented on the basis on technology used.

Heart disease is a leading cause of death worldwide. However, it remains difficult for clinicians to predict heart disease as it is a complex and costly task. Hence, [2] they proposed a clinical support system for predicting heart disease to help clinicians with diagnostic and make better decisions. Machine learning algorithms such as Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Random Forest, and Decision Tree are applied in this study for predicting Heart Disease using risk factors data retrieved from medical files. Several experiments have been conducted to predict HD using the UCI data set, and the outcome reveals that Naïve Bayes outperforms using both cross-validation and train-test split techniques with an accuracy of 82.17%, 84.28%, respectively. The second conclusion is that the accuracy of all algorithm decrease after applying the cross-validation technique. Finally, we suggested multi validation techniques in prospectively collected data towards the approval of the proposed approach.

Detection of heart disease through early-stage symptoms is a great challenge in the current world scenario. If not diagnosed timely then this may become the cause of death. In developing countries where heart specialist doctors are not available in remote, semi-urban, and rural areas; an accurate decision support system can play a vital role in early-stage detection of heart disease. In this paper [3], the authors have proposed a hybrid decision support system that can assist in the early detection of heart disease based on the clinical parameters of the patient. Authors have used multivariate imputation by chained equations algorithm to handle the missing values. A hybridized feature selection algorithm combining the Genetic Algorithm (GA) and recursive feature elimination has been used for the selection of suitable features from the available dataset. Further for pre-processing of data, SMOTE (Synthetic Minority Oversampling Technique) and standard scalar methods have been used. In the last step of the development of the proposed hybrid system, authors have used support vector machine, naive bayes, logistic regression, random forest, and adaboost classifiers. It has been found that the system has given the most accurate results with random forest classifier. The proposed hybrid system was tested in the simulation environment developed using Python.

Heart disease is one of the significant reason of death and disability. The shortage of Doctors, experts and ignoring patient symptoms lead to big challenge that may cause death, disability to the patient. Therefore, we need expert system that serve as an analysis tool to discover hidden information and patterns in heart disease medical data. Data mining is a cognitive procedure of discovering the hidden approach patterns from large data set. The available massive data can used to extract useful information and relate all attributes to make a decision. Various techniques listed and tested here to understand the accuracy level of each. In previous studies, researchers expressed their effort on finding best prediction model. This paper [4] proposes new heart disease prediction system that combine all techniques into one single algorithm, it called hybridization. The result confirm that accurate diagnose can be taken by using a combined model from all techniques.

Heart disease is one of the most common causes of death around the world nowadays. Often, the enormous amount of information is gathered to detect diseases in medical science. All of the information is not useful but vital in taking the correct decision. Thus, it is not always easy to detect the heart disease because it requires skilled knowledge or experiences about heart failure symptoms for an early prediction. Most of the medical dataset are dispersed, widespread and assorted. However, data mining is a robust technique for extracting invisible, predictive and actionable information from the extensive databases. In this paper [5], by using info gain feature selection technique and removing unnecessary features, different classification techniques such that KNN, Decision Tree (ID3), Gaussian Naïve Bayes, Logistic Regression and Random Forest are used on heart disease dataset for better prediction. Different performance measurement factors such as accuracy, ROC curve, precision, recall, sensitivity, specificity, and F1-score are considered to determine the performance of the classification techniques. Among them, Logistic Regression performed better, and the classification accuracy is 92.76%.

Health care industries are one among the top in generating data. To mine the complex data advance algorithms and techniques are needed. The data extraction techniques are used to convert these raw facts as meaningful information. One of the popular data extraction techniques is data mining and machine learning. With the patient data Health care industries are now focusing on optimizing the efficiency and quality of the treatment using various data analytical tools. Data mining and Machine learning has been used by many industries, however they are the proven methodology in health care. Non communicable disease such a heart disease, diabetics and cancer are major reason for the death around the world. Heart disease is one among the top reason for death. In this research paper [6] they have implemented popular data mining algorithms viz., Support vector machine and decision tree with the relevant heart disease data set using Python.

Nowadays, machine learning algorithms have become very important in the medical sector, especially for diagnosing disease from the medical database. Many companies using these techniques for the early prediction of diseases and enhance medical diagnostics. The motivation of this paper [7] is to give an overview of the machine learning algorithms that are applied for the identification and prediction of many diseases such as Naïve Bayes, logistic regression, support vector machine, K-nearest neighbor, K-means clustering, decision tree, and random forest. In this work, many previous studies were reviewed that used machine learning algorithms for detecting various diseases in the medical area in the last three years.

Disease diagnosis is the most critical health-care function. If an illness is diagnosed before the normal or planned period it can save people’s lives. Classification method of machine learning can be useful to help the medical branch by delivering reliable and instant disease diagnosis. Hence the convenient time for both physicians and patients because heart disease is one of the world’s most ultra-hazardous and dangerous diseases today, due to the difficulty to diagnose the disease. Within this paper [8] they include a review of the classification methods for machine learning and image fusion that have been demonstrated to help healthcare professionals identify heart disease.

Cardiovascular diseases (CVD) are among the most common serious illnesses affecting human health. CVDs may be prevented or mitigated by early diagnosis, and this may reduce mortality rates. Identifying risk factors using machine learning models is a promising approach. They would like to propose [9] a model that incorporates different methods to achieve effective prediction of heart disease. For our proposed model to be successful, they have used efficient Data Collection, Data Pre-processing and Data Transformation methods to create accurate information for the training model. They have used a combined dataset (Cleveland, Long Beach VA, Switzerland, Hungarian and Stat log). Suitable features are selected by using the Relief, and Least Absolute Shrinkage and Selection Operator (LASSO) techniques.

Cardiovascular disease has become one of the world's major causes of death. Accurate and timely diagnosis is of crucial importance. They constructed an intelligent diagnostic framework for prediction of heart disease, using the Cleveland Heart disease dataset. They have used three machine learning approaches, Decision Tree (DT), KNearest Neighbor (KNN), and Random Forest (RF) in combination with different sets of features. They have [10] applied the three techniques to the full set of features, to a set of ten features selected by "Pearson's Correlation" technique and to a set of six features selected by the Relief algorithm. Results were evaluated based on accuracy, precision, sensitivity, and several other indices. The best results were obtained with the combination of the RF classifier and the features selected by Relief achieving an accuracy of 98.36%.

IV. Conclusion

The diagnosis of heart disease is usually based on signs, symptoms, and physical examination of the patient. Several factors increase the risk of heart diseases, such as smoking habits, body cholesterol levels, family history of heart disease, obesity, high blood pressure, and lack of physical exercise. However, early diagnosis of heart disease, can help in making decisions on lifestyle changes in high-risk patients and help reduce complications. Here we present different classification techniques for heart disease prediction systems, in future work, we proposed an efficient model with machine learning techniques to improve the result and performance parameters.

References:

[1] Sakshi Goel, Abhinav Deep, Shilpa Srivastava, "Comparative Analysis of various Techniques for Heart Disease Prediction", IEEE, International Conference on Information Systems and Computer Networks, 2019, pp. 88-94.

[2] Halima El Hamdaoui, Saïd Boujraf, Nour El Houda Chaoui, Mustapha Maaroufi, "A Clinical support system for Prediction of Heart Disease using Machine Learning Techniques", International Conference on Advanced Technologies For Signal and Image Processing, 2020, pp. 1-5.

[3] Pooja Rani, Rajneesh Kumar, Nada M. O. Sid Ahmed, Anurag Jain, "A decision support system for heart disease prediction based upon machine learning", Journal of Reliable Intelligent Environments, Springer, 2021, pp. 1-14.

[4] Monther Tarawneh, Ossama Embarak, "Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques", ACTA scientific nutritional health, 2019, pp. 1-6.

[5] S. M. M. Hasan, M. A. Mamun, M. P. Uddin, M. A. Hossain, "Comparative Analysis of Classification Approaches for Heart Disease Prediction", 2020, PP. 1-4.

[6] S.Poonguzhali, P.Sujatha, P.Sripriya, V.Deepa, K. Mahalakshmi, "Performance Evaluation of Classification Methods For Predicting Heart Disease", 2019, pp. 1-9.

[7] Ibrahim Mahmood Ibrahim, Adnan Mohsin Abdulazeez, "The Role of Machine Learning Algorithms for Diagnosing Diseases", Journal of Applied Science And Technology Trends, 2021, pp. 1-10.

[8] Manoj Diwakar, Amrendra Tripathi, Kapil Joshi, Minakshi Memoria, Prabhishek Singh, Neeraj kumar,

“Latest trends on heart disease prediction using machine learning and image fusion”, Materials Today: Proceedings, 2020, pp. 1-7.

[9] Pronab Ghosh, Sami Azam, Mirjam Jonkman, Asif Karim, “Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques”, IEEE Access, 2021, pp. 19304-19326.

[10] Pronab ghosh, Sami Azam, Asif Karim, “Use of Efficient Machine Learning Techniques in the Identification of Patients with Heart Diseases”, ICISDM 2021, pp. 14-20.

[11] P Kalpana, S Shiyam Vignesh, L M P Surya, V Vishnu Prasad, “Prediction of Heart Disease Using Machine Learning”, Journal of Physics: Conference Series, 2021, pp. 1-5.

[12] Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, Preeti Nagrath, “Heart disease prediction using machine learning algorithms”, IOP Conf. Series: Materials Science and Engineering, 2021, pp. 1-11.

[13] Abid Ishaq, Saima Sadiq, “Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques”, IEEE Access, 2021, pp. 39707-39716.