IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Disease Prediction Using Data Mining Techniques – A Survey

Ovias Tajdar[1*], Bhavya Alankar[2]

[1, 2] School of Engineering Science and Technology (SEST), Jamia Hamdard, New Delhi, India

Corresponding Author: oviasdar@gmail.com, Tel. 7006938151

DOI: https://doi.org/10.26438/ijcse/v7i4.10701075 | Available online at: www.ijcseonline.org

Accepted: 20/Apr/2019, Published: 30/Apr/2019

***Abstract*—** The healthcare industry generates huge data that cannot be handled manually. Using data mining methods, valuable information is extracted from this data to create a relationship between attributes. Machine learning algorithms and data mining techniques are used from data sets to predict the disease. Data mining techniques are used to study disease occurrence. One of the most frequently encountered problems in medical centres is that not all specialists are equally qualified and can give their own conclusion, which can cause the patient to die. Data mining techniques and machine learning algorithms play a dynamic role in the automatic diagnosis of diseases in health care centres to overcome such glitches prediction of diseases. The purpose of this survey paper is to analyse the prior health care research work and advanced disease analysis technologies. Support Vector Machine, Decision Tree, Naïve Bayes, K-Nearest Neighbour, and Artificial Neural Network are some machine algorithms used to predict the occurrence of diseases. Our study concludes that Support Vector Machine shows approximately 85% accuracy and has the potential to be considered as one of the disease prediction capable algorithms.

***Keywords*—** Data mining , Machine Learning, Support Vector Machine, Decision Tree, Naïve Bayes, K-Nearest Neighbor, and Artificial Neural Network

I. INTRODUCTION

A foremost task facing health care organizations is the delivery of good quality services at reasonable budgets. The Medical centres are duty-bound to be able to diagnose patients correctly Poor clinical conclusions can lead to tragic significances which must be avoided. Hospitals can reach to better conclusions by hiring suitable computer based information and decision support systems. These systems normally produce massive amounts of data which remains largely idle and needs to be converted into useful data. Data Mining is the removal of hidden, anonymous and likely valuable information about data. It is the practice of analysing data and collecting knowledge from it. It is critical in many fields of studies to determine hidden information from gigantic datasets that helps data scientists to recognize and recover their data within a short period. Data mining techniques are used to categorize, forecast and cluster data to make accurate decision making in many organizations. The consequences of these systems are to deliver assistances to health care organizations for clustering the patients having similar types of health issues so that health care organization offers their actual treatments. Data Mining and its applications in public health is a new field of study.

Prediction is a worthy practice in health care centres where clinicians can make improved decisions and avoid the situations that may lead to the death of patients which is quite unacceptable. Data mining gives several models for diagnosis of diseases such as supervised, unsupervised, ensemble and hybrid classification. Classification is defined as supervised learning which classifies each data item into one of the predefined classes or groups. Clustering which is unsupervised learning technique clusters data into groups of similar objects which have the same properties, and distinguishes from the objects having dissimilar properties forming another group. Various merging methods are well-defined under ensemble learning and intend to realize improved accuracy over single classifier models. Combining heterogeneous learning approaches comes under hybrid learning methods. Different data mining techniques are reviewed and presented in this paper. Subsequent sections describe the most debated data mining techniques for prediction of disease. Section I contains the introduction of data mining and its applications, Section II contains the tools of data mining , Section III discusses the literature survey and Section IV concludes research work with future directions.

A. Artificial Neural Network

An Artificial Neural Network (ANN) also termed as Neural Network is a mathematical model that emulates a biological neural system. A multi-perceptron neural network is frequently used. It plots a set of input data onto a set of

correct output data. It consists of three layers-input layers, an output layer, and hidden layer. There is an edge between each layer and weights are allotted to each edge. The key purpose of neurons of the input layer is to split the input into neurons in the hidden layer. Neurons of the hidden layer enhance input signal with weights of edges from the input layer. The output of the Neural Network uses a function such as sigmoid function or hyperbolic tangent function.

B. Support Vector Machine

The notion of Support Vector Machine is given by Vapnik., which is centred on statistical learning theory. SVMs were originally developed for binary classification but it could be competently extended for multiclass problems. The support vector machine classifier creates a hyperplane or multiple hyperplanes in high dimensional space that is favourable for classification, regression, and other operative tasks. SVM has many striking features. It constructs a hyperplane in original input space to disperse the data points. Kernel functions are used for non-linear mapping of training samples to higher dimensional space. Numerous kernel function such as polynomial, Gaussian, sigmoid etc., are used for this purpose. SVM works on the idea that data points are classified using a hyperplane which maximizes the separation between data points and the hyperplane is built with the help of support vectors.

C. Decision Trees

The decision tree methodology is more influential for classification problems. There are two stages in this technique, i.e. constructing a tree & relating the tree to the dataset. There are many prevalent decision tree algorithms CART, C4.5, J48, ID3, and CHAID. From these, J48 algorithm is used for this scheme. J48 algorithm uses a pruning method to construct a tree. Pruning is a technique that shrinks the size of the tree by eliminating overfitting data, which leads to reduced accuracy in predictions. The J48 algorithm recursively organizes data until it has been classified as seamlessly as possible. This technique gives better accuracy results on training data. The broad concept of this algorithm is to build a tree that offers the stability of flexibility & precision.

D. Naïve Bayes

Naive Bayes classifier is grounded on Bayes theorem which uses conditional independence i.e. it assumes that any particular feature in the class is not linked to existence of any feature. The Bayes theorem follows as: Let $X=\{x_1, x_2, ...,x_n\}$ be a set of n attributes. In Bayesian network, X is measured as evidence and H denotes some hypothesis means, the data of X belongs to specific class S. We have to define P (H|X), the probability that the hypothesis H holds with supposed evidence i.e. data sample X. According to Bayes theorem the P (H|X) is defined as P (H|X) = P (X| H) P (H) / P (X),where P(H/X) is Posterior Probability, P(X/H) is likelihood, P(X) is the class prior probability and P(X) is predicted prior probability.

II. TOOLS AVAILABLE FOR DATA PREPROCESSING AND CLASSIFICATION

There are numerous data mining tools which can be used for data pre-processing and classification with the help of machine learning techniques. Some influential tools available for data mining are discussed.

1) Rapid Miner: Rapid Miner is a data science tool and is written in JAVA language. It offers an integrated platform for deep learning, predictive analytics and machine learning. This tool can be used for applications ranging from businesses to machine learning.

2) Orange: Orange is the software platform for machine learning and data mining. It can be best used for data visualization and is written in the Python language. Orange is fairly great to control and is more interactive.

3) WEKA: This software is also known as the Waikato Environment. It is suitable for data analysis and predictive modelling. It comprises of algorithms and tools that provision machine learning and is written in JAVA language.

4) KNIME: KNIME is one of the integration platforms for data analysis. It uses the concept of the modular data pipeline. KNIME is known for some well-known features like quick deployment. Users learn to handle KNIME very quickly.

5) Oracle data mining: Oracle data mining delivers exceptional algorithms for classification, regression, and prediction. It is known for making better predictions, target best customers and detects fraud. It offers the capability of "drag and drop" of data inside the database and provides better understandings.

6) Rattle: Rattle is a Graphical User Interface based tool that uses R as underlying programming language and is used in many situations. Rattle provides substantial data mining functions. The data created by Rattle can be observed and corrected. The code can be studied and extended without any limitation.

7) DataMelt: DMelt is written in JAVA and is a computational platform. It can be used for study of massive data. It can be used on multiple operating systems which can be executed with Java Virtual Machine.

III. LITERATURE SURVEY

Abundant research has been done that has focused on the diagnosis of diseases especially Coronary Heart Disease (CHD), prediction of brain tumor, bone tumor, diabetes mellitus. They have used different data mining methods for diagnosis and calculated different probabilities for different methods.

Paper [1] classifies heartbeat time series using a support vector machine. SVM classifier is compared to other neural network based classification techniques. The author is of the opinion the main characteristic of already proposed methods is that they focus on a specific feature in order to improve performance. Feature selection is an important topic discussed here and analysis is done which confirm the effectiveness of SVM even in the presence of some amount of noise. Research conducted in this paper compared SVM classification of heart rate signals with the classifications acquired by other nonlinear classifiers also confirmed the efficacy of the first methodology even in the occurrence of noise.

Kurt et al [2] compares the performances of classification techniques in order to predict the presence of Coronary Artery Disease (CAD). Logistic Regression (LR) is compared with classification and Regression Tree (CART), multi-layer perceptron (MLP), radial basis function (RBF), and self-organizing feature maps (SOFM). This paper compares methods by using a real data set in order to provide general information of data structures that help researchers to select the best method for solving problems based on classification. Authors suggest that data should be explored and processed by high-performance modelling methods. Areas under ROC curve were calculated as 0.783, 0.753, 0.745, 0.721 and 0.625 for MLP, LR, CART, RBF and SOFM. MLP was found to be the best procedure to predict the existence of CAD in the data set.

Paper [3] compares the performance of six classifiers in the prediction of chronic kidney disease. The results show that the Random Forest (RF) classifier outperforms Sequential Minimal Optimizations (SMO), Naive Bayes, Radial Basis function (RBF), Multi-Layer Perceptron and Simple Logistic in terms of Area under the ROC curve and accuracy. It was also detected that few classifiers have produced poor classification accuracy as compared to RF like SMO and RBF.

Paper [4] studies diabetes recognition using different machine learning techniques based on tenfold cross-validation. Logistic regression outperformed all other algorithms used in the paper like Naïve Bayes, Classification Tree, Support Vector Machine, K-Nearest Neighbors, and Artificial Neural Network. This paper has involved multiple classification approaches in the identification of diabetes centred on many clinical parameters. These approaches were analysed in terms of eight parameters. However, Logistic Regression gives better classification results with the highest accuracy of 78% whereas Artificial Neural Network performed at 77% accuracy.

Paper [5] introduces Bayesian inference as a decision-making tool to guide radiologists for correct diagnosis of the bone tumour since this tumour comes in different forms. Naïve Bayes machine is built in such a way that performs Leave-one-out cross-validation. Primary accuracy (PA) was 44% and differential accuracy (DA) was 60% for 29 common diagnoses (710 cases) whereas PA was calculated as 62% and DA was calculated to 80% for 10 most common diagnoses (478 cases). The proposed model can be used as an alternative to SVM or deep neural networks that require large training data.

Hlaudi et al [6] conducted research based on the experiment of various data mining algorithms to predict heart attacks. The paper studies various parameters and compares the best method for prediction. Data mining algorithms such as Naive Bayes, CART, REPTREE, Bayes Net, and J48 are used to predict heart attacks. The results show all algorithms performed better with prediction accuracies of 99%.The experiment suggested in this paper can be used as an important device for clinicians to predict hazardous cases and counsel consequently. This model will be able to answer difficult queries in the prediction of heart attack diseases.

Authors in paper [7] explained the importance of mining huge amount of data to discover hidden information for decision making. This paper introduced Intelligent Heart Disease Prediction system (IHDPS) as a prototype which uses data mining techniques namely Neural Network, Decision trees, and Naive Bayes. The prototype is web-based, scalable, user-friendly, and implemented on the .NET platform. IHDPS can assist as a teaching tool to help nurses and medical scholars to analyse the patients with heart diseases. It can also provide decision support to help doctors to make improved clinical decisions .The results show that each technique has its own strength to get suitable results.

Paper [8] analysed the performances for ten classification data mining techniques on a real-time patient database and a confusion matrix was displayed for fast check. PLS-DA showed solid results than any other classifier and paper suggested using this classifier to get better results with accuracy and performance. Tanagra is a data mining matching set used in this paper and gives the finest results.

Authors predicted heart disease, Blood Pressure and sugar with neural networks in paper [9] and suggested supervised network for heart disease diagnosis and trained it using Back Propagation Algorithm. When unknown data is entered by the doctor, the system will use trained data and find the unknown data and produce a list of possible diseases patient may be suffering from. The system introduced is more reliable and efficient. It is concluded from this research work that ANN is an alternative to classical statistical techniques for modelling and predicting data.

Qurat et al. [10] describes the automatic detection of Coronary Artery Disease (CAD). This paper identifies the best methods and classifier for identification of CAD. Two workflows are proposed which are necessary to follow for future detection of CAD at an early stage in order to reduce cost. SVM classifier represents a favourable approach for CAD identification which gives an accuracy of 99.2%. K-fold cross-validation is measured here. The authors suggest that the performance of the classifier lies in the nature and size of the data set.

Lahsasna et. al. [11] presents a Fuzzy Rule-Based System (FRBS) to aid as a decision support system for Coronary Heart Disease (CHD) diagnosis that considers accuracy and transparency at the same time. Ensemble Classifiers Strategy (ECS) method is projected to enhance the classification ability of FRBS. Results reveal that the accuracy of generated rules is better than traditional classification methods. The relations between factors and CHD diagnosis may reveal unpredicted findings and knowledge that can be used to detect CHD existence at an early stage and may result in major life-saving.

Authors in [12] investigated heart disease prediction using data mining techniques – K-star, J48, Bayes Net, SMO, and Multi-Perceptron through WEKA software. Two data sets are used separately. Based on performances, SMO and Bayes Net achieve optimum performance (89% and 87% respectively) than K Star, Multi-layer Perceptron and J48 techniques. K-fold cross-validation is considered.

Paper [13] intends at predicting the conclusion of stroke using KDP methods, ANN and SVM models. KDP is used for discovering patterns and knowledge in the given data. KDP was used for better understanding of data, reduced cost, time and unnecessary task to be performed. Pattern extraction was done but ANN and SVM. ANN model had a better predictive performance for stroke as compared with SVM. Accuracy value for ANN was 81.82% while as Accuracy value for SVM was 80.38%

The authors in paper [14] provide an investigation of different papers for the prediction of heart disease considering the performance of data mining algorithms. This paper gives the information regarding different data mining tools and their importance in real world. The survey shows that SVM ascertains to be efficient and effective in terms of accuracy as compared to other data mining techniques in the prediction of heart disease. This paper shows the result of different technologies and their accuracy with respect to each other.

In paper [15], the researchers compared the accuracy of Multi-layer Perceptron (MLP) Neural Network and Support Vector Machine (SVM) on heart Disease data set. The data set was used in ARFF (Attribute Relation File Format) format, supported by WEKA machine software and implemented the classification mentioned in the paper. SVM proves to be efficient than Multi-Layer Perceptron Neural Network using a dataset of 303 patients for classification. SVM achieved 84.4884% accuracy while as MLP was 80.5281% accurate.

Paper [16] predicts heart disease more accurately. Researchers used two more input attributes, obesity and smoking to get better results. Three data mining technologies were applied namely Decision Trees, Naïve Bayes, and Neural Network. From results, the neural network provides accurate results compared to Decision Trees and Naïve Bayes, Neural Networks. Accuracy with Neural networks was calculated as 100% considering 15 attributes.

P. Kumar et al. [17] recognized clustering data mining as a classification procedure for the inspection of K- means and X-means algorithm. The paper studies tumour analysis. These data mining algorithms were used to organize colon data set into two classes. The paper concludes that the accuracy attained by k-means is more than x-means algorithm while as x-means algorithm performs better in case of execution speed.

Emrana Kabir Hashi et al. [18] described an expert clinical decision support system to predict the disease using classification methods. This system uses insight and knowledge of doctor without using complex clinical data. The suggested system supports the doctor to make prediction and also benefits the patients as well as medical insurance companies. The paper is based on WEKA software and percentage ratio method for train and test dataset were calculated using C4.5 and KNN which gave 90.43% and 76.96% accuracy respectively. C4.5 Decision Tree gave better accuracy compared to KNN.

Paper [20] briefly examines the potential utilization of classification-based data mining techniques such as Rule-based, Decision Tree, Naïve Bayes, and Artificial Neural Network for massive volume of health care data. In this paper, authors used data mining to present an intelligent and effective method of heart attack prediction. First, they provided an efficient approach for extracting significant patterns from cardiac data warehouses for the efficient prediction of cardiac attack Based on the calculated significant weight; frequent patterns with a value greater than a predefined threshold were chosen for the valuable prediction of cardiac attack.

Paper [21] summarized various reviews and technical papers on the diagnosis and prognosis of breast cancer. In this paper authors have presented an overview of the current research

being conducted using the techniques of data mining to improve the diagnosis and prognosis of breast cancer.

Table 1 Comparison of different data mining algorithms on basis of accuracy

Author	Purpose	Techniques used	Accuracy
Manish et. al.[5]	Prediction of chronic kidney disease	Random Forest	100%
		Naïve Bayes	95%
		Sequential Minimal Optimization	97.8%
		Radial basis function	98.8%
		Multi-layer Perceptron	98%
		Simple Logistic	98%
Bao.H.Do et.al [6]	Bone Tumor Diagnosis using NB model	NB primary Accuracy	62%
		NB differential accuracy	80%
K.R.Lakshmi[8]	Prediction of Heart Disease Survivability	C4.5	84.68%
		BLR	82.96%
		PLS-DA	86.13%
		EMC	85.16%
Marjia et.al [12]	Heart Disease Prediction	K-Star	75%
		J48	86%
		SMO	89%
		Bayes Net	87%
		Multi-layer Perceptron	86%
Parisa et. al.[15]	Classification of Health care data using SVM and MLP	MLP	80.5281%
		SVM	84.4884%
Parvesh et.al [17]	Tumour classification	Global K-means	59.68%
		X-means	58.06%
A.K.Dwivedi[4]	Diabetes Mellitus Prediction	SVM	82%
		ANN	84%
		Logistic Regression	85%
		KNN	80%
		Classification	77%
		Naïve Bayes	83%
Emrana Kabir Hashi [18]	Predict disease using classification technique	C4.5	90.43%
		KNN	76.96%

Paper [19] gives overview of data mining techniques to identify liver disease at early stage. This paper studies algorithms such as C4.5, Naïve Bayes, Decision Tress, Support Vector Machine (SVM), Back Propagation Neural Network (BPNN), and Classification and Regression Tree (CART).These algorithms give different results in terms of speed, accuracy, cost and performance and cost. This paper lists the advantages and disadvantages of each algorithm. It concludes that C4.5 is used to handle discrete and continuous values and gives better results as compared to other algorithms.

IV. CONCLUSION AND FUTURE SCOPE

This paper carries out a comprehensive analysis of the various data mining technologies available and to ascertain the best among them. In this survey, we found that there is still room for improvement in classification and prediction of disease. One of the benefits of survey papers is to recover the existing methodology for improved decision-making process by using various algorithms and feature extraction algorithm. The overall objective of this paper was to investigate various data mining techniques which have emerged recently to predict heart disease and to compare them to find out the best method for prediction. The analysis shows that different technology uses a different number of attributes and show different accuracy to each other. SVM proves to have about 85% accuracy and has the potential to be considered one among the capable algorithms in the prediction of disease. Many authors suggest that the performance of the algorithm lies in the nature and accuracy of the data set. The application of data mining in healthcare for diagnosis of disease is increasing manifold and as already stated, the prediction accuracy of existing systems can be enhanced, so in future, there is a scope for new algorithms which overcome the drawbacks of the existing system.

REFERENCES

[1] Kampouraki, A., Manis, G., & Nikou, C. "*Heartbeat Time Series Classification with Support Vector Machines*," IEEE Transactions on Information Technology in Biomedicine, 13(4), 512–518, 2009.

[2] Kurt, I., Ture, M., & Kurum, A. T. "*Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease*," Expert Systems with Applications, 34(1), 366–374, 2008.

[3] Manish Kumar, "*Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm*," International Journal of Computer Science and Mobile Computing, Vol.5 Issue.2, pg. 24-33, Feb.2016.

[4] Dwivedi, A. K. "*Analysis of computational intelligence techniques for diabetes mellitus prediction*," Neural Computing and Applications, 2017.

[5] Do, B. H., Langlotz, C., & Beaulieu, C. F. "*Bone Tumor Diagnosis Using a Naïve Bayesian Model of Demographic and Radiographic Features*," Journal of Digital Imaging, 30(5), 640–647, 2017.

[6] Hlaudi Daniel Masethe, Mosima Anna Masethe, "*Prediction of Heart Disease using Classification algorithms*". Proceedings of the World Congress on Engineering and Computer Science 2014 Vol. II,WCECS San Francisco, USA, 22-24 October, 2014

[7] Sellappan Palaniappan, Rafiah Awang, "*Intelligent Heart Disease Prediction System Using Data Mining Techniques*," IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008

[8] K.R. Lakshmi , M.Veera Krishna and S.Prem Kumar, "*Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability* ",International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013.

[9] Carlos Ordonez, "*Improving Heart Disease Prediction Using Constrained Association Rules*," Seminar Presentation at University of Tokyo, 2004.

[10] Mastoi, Q., Wah, T. Y., Gopal Raj, R., & Iqbal, U. "*Automated Diagnosis of Coronary Artery Disease: A Review and Workflow*". Cardiology Research and Practice, 2018

[11] Lahsasna, A., Ainon, R. N., Zainuddin, R., & Bulgiba, A. "*Design of a Fuzzy-based Decision Support System for Coronary Heart Disease Diagnosis*". Journal of Medical Systems, 36(5), 2012

[12] Sultana, M., Haider, A., & Uddin, M. S. "*Analysis of data mining techniques for heart disease prediction*," 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 2016.

[13] Colak, C., Karaman, E., & Turtay, M. G. "*Application of knowledge discovery process on the prediction of stroke*," Computer Methods and Programs in Biomedicine, 119(3), 181–185, 2015.

[14] Megha Shahi1, Er. Rupinder Kaur Gurm., "*Heart Disease Prediction System Using Data Mining Techniques - A Review*," International Journal of Technology and Computing (IJTC) ISSN-2455-099X, Volume 3, Issue 4, April 2017.

[15] Naraei, P., Abhari, A., & Sadeghian, A. "*Application of multilayer perceptron neural networks and support vector machines in classification of healthcare data*". Future Technologies Conference (FTC), 2016.

[16] M. Shahi and R. Kaur Gurm, "*Heart disease prediction system using data mining techniques*," Orient. J. Computer Science Technology, vol. 6, no. 4, pp. 457–466, 2013.

[17] Kumar, P., & Wasan, S. K. "*Analysis of X-means and global k-means using tumor classification*," The 2nd International Conference on Computer and Automation Engineering (ICCAE), 2010.

[18] Hashi, E. K., Zaman, M. S. U., & Hasan, M. R. "*An expert clinical decision support system to predict disease using classification techniques*," International Conference on Electrical, Computer and Communication Engineering (ECCE), 2017.

[19] D.Sindhuja,R. Jemina Priadarsini, "*A Survey on Classification Techniques in Data Mining for Analyzing Liver Disease Disorder*," International Journal of Computer Science and Mobile Computing. Vol.5 Issue.5, pg. 483-488, May 2016.

[20] K.Srinivas B.Kavihta Rani Dr. A.Govrdhan. "*Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks*," International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 250-255

[21] Shelly Gupta, Dharminder Kumar, Anand Sharma, "*Data mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis*", Indian Journal of Computer Science and Engineering (IJCSE).

Authors Profile

Ovias Tajdar completed Bachelor of Technology from Baba Ghulam Shah Budshah University, Rajouri, Jammu and Kashmir in 2012 and is currently pusuing Master of Technology from Jamia Hamdard, New Delhi. His main research work focuses on Data Mining, Big Data Analytics, Artificial Intelligence and Machine Learning.

Dr. Bhavya Alankar completed BTech from Utthrakhand Technical University and Mtech from CDAC, Mohali. He has done his PhD from Uttrakhand Technical Univeristy. He has 12 years of teaching experience.

