A Review Study on Data Mining Algorithms for Prediction Diseases

[1]Er. Manpreet Kaur, Master of Engineering Student, Guru Nanak Dev Engineering College, Ludhiana, India

[2]Er. Shailja, Assistant Professor, Guru Nanak Dev Engineering College, Ludhiana, India.

Abstract: The healthcare industry assembles massive volume of healthcare information or data that circulate the information into useful data. In everyday life several factors that affect the human diseases. Hospitals are producing large amount of information related to patients. This paper describes the various data mining algorithms such as neural network, support vector machine, KNN, decision tree etc. and provides an overall brief of the existing work. The major advantage of using data mining is that to identify the structures.

Keywords —Data mining, Classification, Clustering, Prediction, Model selection, Analysis model

I. INTRODUCTION

Data Mining means withdrawal of knowledgeable information from a large data. Data mining is not only withdrawal of knowledgeable information, it also mine the data in the form of patterns and figures. Data mining is also known as KDD process. It follows mainly three factors:

- Data pre-processing
- Data extraction
- Data presentation

Data mining can be used in different area such as detection, marketing, healthcare, management etc. It also helps to predict the diseases like heart attack, lung cancer, Brest cancer, diabetes prediction etc. using patient's database. Classification and regression both techniques also used for disease prediction. Heart is the fundamental part of our body. Heart diseases mostly occur due to poor healthy lifestyle, smoking, alcohol which may inducement. hypertension, high blood sugar, high blood pressure. Data mining can help to decrease the risk of heart disease by techniques. A major objective covering healthcare zone is quality of supply. Quality of supply include analyzing the diseases correctly and implement the medical treatments to patients. According to sample of world health organization 17 million people are died by diseases. Various techniques are help for prediction such as random forest, k nearest neigbours, support vector machine, neural network, density-based clustering, naïve Bayes, decision tree etc.

Classification Techniques: - The classification technique in data mining to take a dataset from hospitals and assign the desired class to dataset. Classification method is used to reduce the number of FN and FP that is false negative and false positive. Some algorithms like Naïve bayes, decision tree, KNN, Support vector machine etc are covers under classification techniques.

Clustering Techniques: - Clustering is a mechanism that describes the pattern of dataset by dividing it into proper clusters. The word 'clustering' helps to identify structure for distribution of enabled data. It is a unsupervised learning system that provides high preferable cluster with huge integrals similarity and low integrals similarity. Basically, cluster is a company of objects that linked to the same class. It also helps to define on arrangement of structures in one form by distance function. Clustering has two types are- hard and soft clustering. Various techniques are used in clustering: -

K-mean clustering

Hierarchical clustering

Hard clustering contains data point that either linked to the cluster or not. Soft clustering measures the probability of data point to be cluster is appointed and each data point is lie on separate cluster.

II. LITERATURE SURVEY

Moloud Abdar et al., conducts a data mining technique to predict the risk of heart diseases. After feature analysis five algorithms including C5.0, Neural Network, Support Vector Machine (SVM), K-Nearest Neighborhood (KNN) and Logistic Regression, developed and validated. C5.0 Decision tree has been able to build a model with greatest accuracy 93.02%, KNN, SVM, Neural Network have been 88.37%, 86.05% and 80.23% respectively[1].

Kalyani A.Bhawar et al., presents an algorithm about brain tumor classification is to calculate vector patterns and classify the tumors automatically. The purpose of this work is to present an updated survey of current methods for

 DOI : 10.35291/2454-9150.2020.0340 © 2020, IJREAM All Rights Reserved.

making decision tree for classifying brain tumors. The main goal is on solving the cancer classification problem using single decision tree classifiers[9].

Irina Ionita et al., presented four classification models: Naive Bayes, Decision Tree, Multilayer Perceptron and Radial Basis Function Network. The results produce an accuracy for all the classification models and the best classification rate acquired from the Decision Tree model. The data set used to validate the classifier was provided by UCI machine learning repository. The framework for building and testing the classification models was Weka, data mining software[6].

K.Gomathi et al., present an analysis of the Heart disease for male patients using data mining techniques. The

Pre-processed data set consists of 210 records in which available 8 attributes from the database. We have studied three data mining techniques: Naïve Bayes, Artificial neural network, and the J48 decision tree algorithms. Our result Shows that of these three classification models Naïve Bayes predicts heart disease with higher Accuracy[12].

Purushottam et al., proposed a framework that can find the risk level of patients in view of the given parameter about their health. The main objective of this study is to help a doctor to make correct decision about the heart disease risk level. The rules generated by the proposed system are important as Original Rules, Pruned Rules, rules without duplicates, Classified Rules, Sorted Rules and Polish. The execution of the framework is worked as the results explains that the framework has reduce the coronary illness risk level[10].

Varun Jain et al., analyze different data mining classification methods: Naïve Bayesian, Decision tree, Support Vector Machine and Artificial Neural Network on Brain Tumor data set. These techniques are compared in the terms of sensitivity, specificity, accuracy, error rate, True Positive Rate and False Positive Rate. Our results for Naïve Bayes, Decision Tree, ANN and SVM of accuracy are 92.03%, 93.51%, 91.44%, 93.80% respectively. SVM is the best classifier to detect brain tumor disease with high accuracy and lowest error rate[3].

In this paper we have discussed various data mining approaches that have been utilized for breast cancer diagnosis and prognosis. We have proposed a breast cancer prediction framework consisting of four main modules: Data Collection, Data Preprocessing, Feature Selection, and Classification. Evaluation results are provided as well. The goal is to find the best combination for feature extraction algorithm and classification algorithm, which will improve the accuracy of mammograms classification process.

Luzana Subhani et al., discussed various data mining approaches that have been predicted for breast cancer diagnosis. The framework consisting of four main modules: Data Collection, Data Preprocessing, Feature Selection and Classification. The objective is to find the best combination for feature extraction algorithm and classification algorithm, which will improve the accuracy[7].

Vikas Chaurasia et al., present a report on breast cancer to develop prediction models for breast cancer. We used three popular data mining algorithms (Naïve Bayes, RBF Network, J48) to develop the prediction models using a dataset contains 683 patients We also used 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes. The results indicated that the Naïve Bayes is the best predictor with 97.36% accuracy on the holdout sample[11].

Yugai li et al., proposed use of machine learning methods that combine feature selection to classify and predict diabetes. Feature selection used as input variables of support vector machine (SVM), decision tree, and integrated learning model (Adaboost and Bagging) for modeling and prediction. The results show that Adaboost algorithm produces better classification results[4].

J.Steffi et al., explores the prediction of diabetes using data mining techniques. The dataset has taken 768 instances from Dataset to determine the accuracy of the data mining techniques in prediction. This study is to compare the performance analysis of Naïve Bayes, Logistic Regression, Artificial neural networks (ANNs), C5.0 Decision Tree and Support Vector Machine (SVM) models for predicting diabetes. The decision tree model (C5.0) had given the best classification accuracy[13].

Chaitrali S.Dangare et al., analysed prediction systems for Heart disease using more number of input attributes. The system uses 13 attributes to predict the Heart disease. Until now, 13 attributes are used for prediction. This paper added two more attributes that is obesity and smoking. The data mining classification techniques such as Decision Trees, Naive Bayes, and Neural Networks are analyzed on Heart disease database. Neural Networks predicts Heart disease with highest accuracy[14].

Saurabh Pal et al., conducts two main techniques are presented for mining the hidden pattern in the dataset. Boosting and Bagging both are machine learning techniques. Boosting produced from decision tree and neural network and Bagging produced from combinations of Adaboost and stacking techniques. Finally proposed experiment is conducted by Bagging and boosting[8].

V. Krishnaiah et al., examine the classification-based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network. Using lung cancer symptoms such as age, sex, Wheezing, Shortness of breath,

 DOI : 10.35291/2454-9150.2020.0340 © 2020, IJREAM All Rights Reserved.

Pain in shoulder, chest, arm, it can predict the patients getting a lung cancer disease. The objective of this paper is to propose a model to correct diagnosis of the disease that help the doctor in saving the life of the patient[5].

E.Yatish Vankata Chnadra et al., approaches for treating lung cancer. To predict the patient's data mining techniques can be used with selection of algorithms. The algorithms used to detect the lung cancer are Support vector machine (SVM), Decision tree, k-Nearest neighbour, Random forest, Logistic regression. In this paper two different datasets are compared and on implementation found algorithms have more accuracy on data sets for prediction rate of lung cancer[2].

Table 1: - Summary of various paper related diseases

Year	Title	Authors	Tools used	Techniques used	Accuracy (%)
2013	Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques	V.Krishnaiah Dr.G.Narsimha Dr.N.Subhash Chandra[5]	Weka	Naïve bayes OBANB	84.76 80.60
2015	Comparing performance of data mining algorithms in prediction of heart diseases.	Moloud Abdar Sharareh R.Niakan Kalhori Tole Sutikno[1]	Weka	C5.0 SVM KNN Neural network	87.50 82.6 86.41 85.87
2016	Brain tumor classification using data mining algorithms	Kalyani A.Bhawar Ajay S.Chhajad[9]	Matlab	Cart Radom forest	97.36 98.5
2016	Heart disease prediction using data mining classification	K. Gomathi Dr.Shanmuapriyara[12]	Weka	Naïve bayes ANN Decision tree(J48)	79.90 76.55 77.03
2016	Prediction of thyroid disease using data mining techniques	Irina Ionita Livieu Ionita[6]	Weka	Cart J48 MLP RBF Naïve bayes	89.58 89.68 77.08 79.16 70.83
2016	Efficient heart disease prediction system	Purushottam Kanak Saxene Richa sharma[10]	Weka	SVM NN C4.5 MLP hybrid approach	70.59 76.47 73.53 74.85 86.7
2017	Comparative study of data mining classification methods in brain tumor diseases prediction	Varun jain Sunila Godara[3]	Weka 3.6	Decision tree Naïve bayes ANN SVM	93.51 92.03 91.44 93.80
2017	Enhancing breast cancer detection using data mining classification techniques	Luzana Subhani Bujar Raufi4 Jaumin ajdari[7]	R Studio	SVM Neural Network	83.96 86.93
2018	Prediction of breast cancer using data mining techniques	Vikas Chaurasia Saurabh Pal BB Tiwari[11]	Weka	Naïve bayes J48 RBF network	97.36 96.77 93.41
2018	Analysis and study of diabetes follow up data using a data mining-based approach	Yukai Li Huling Li Hua Yao[4]	Matlab	Decision tree SVM Adaboost Bagging	91.15 92.62 94.84 91.15
2018	Predicting diabetes mellitus using data mining techniques	J. Steffi Dr.R.Balasubramanian Mr.K.Aravind kumar[13]	R Studio	Naïve bayes C5.0 ANN SVM Logistic regression	73.57 74.63 72.29 72.17 74.67
2019	Improved study of heart disease prediction system using data mining classification techniques	Chaitrali S.Dangare Sulabha S.Apte[14]	Weka	Naïve bayes Decision tree Neural network	90.74 96.62 99.25
2019	To generate an ensemble model for women thyroid prediction using data mining techniques	Dhyan Chandra Yadav Saurabh pal[8]	R studio	Bagging Boosting Adaboost Stacking	95.98 98.79 96.39 97.1
2019	Lung Cancer prediction using data mining techniques	E.Yatish Venkata Chandra K.Ravi Teja[2]	Visual Studio	KNN Logistic regression Decision tree SVM Random forest	76.52 92.72 100 63.0 95.8

 DOI : 10.35291/2454-9150.2020.0340 © 2020, IJREAM All Rights Reserved.

III. CONCLUSION

The main focus of this paper for prediction of diseases using data mining tools and algorithms. In conclusion, the main aim of this literature survey to analyze the various data mining algorithms to gives a performance with accuracy for data mining in WHO. Data collection is drifting out using various origin that are initial factors responsible for any character of diseases and using a pattern the data is manufactured.

REFERENCES

[1] Moloud Abdar, Sharareh R.Niakan Kalhoru andTole Sutikno, "Comparing performance of data mining algorithms in prediction of heart diseases." in *International journal of electrical and computer engineering(IJECE)*, vol.5, pp.1569-1576, 2015

[2] E. Yatish Venkata Chandra and K.Ravi Teja, "Lung Cancer prediction using data mining techniques" in *International Journal of Recent Technology and Engineering (IJRTE).* Vol.8 Issue-4, November 2019

[3] Varun jain and Sunila Godara, "Comparative study of data mining classification methods in brain tumor diseases prediction," in *IJCSC*, vol. 8(2), pp. 12-17 June 2017.

[4] Hua Yao, Yukai Li and Huling Li, "Analysis and study of diabetes follow up data using a data mining-based approach," in *Soft computing for analysis of biomedical data*, July 2018.

[5] V.Krishnaiah, Dr.G.Narsimha and Dr.N.Subhash Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques," in *International journal of computer science and technologies(IJCSIT)*, vol. 4(1), 2013, pp. 39-45.

[6] Livieu Ionita and Irina Ionita, "Prediction of thyroid disease using data mining techniques," in *Board research in artificial intelligence and neuroscience (BRAIN)*, vol.7, august 2016.

[7] Luzana Subhani, Bujar Raufi and Jaumin ajdari, "Enhancing breast cancer detection using data mining classification techniques," in *IJECE*, 2017.

[8] Dhyan Chandra Yadav and Saurabh pal, "To generate an ensemble model for women thyroid prediction using data mining techniques" in *Asian pacific journal of cancer prevention (APOCP)*, 2019

[9] Kalyani A.Bhawar and Ajay S.Chhajad, "Brain tumor classification using data mining algorithms," in *International journal of engineering sciences and research technology*, vol.5(11), November 2016

[10] G. Purusothaman, Kanak Sexana and Richa sharma, "Efficient heart disease prediction system," in Procedia Computer Science (Elesvier), 2016.

[11] Vikas Chaurasia, Saurabh pal and BB Tiwari, "Prediction of breast cancer using data mining techniques", in *International journal of computer science and mobile computing*, vol.3(1), pp. 10-22, January 2016.

[12] K. Gomathi and Dr.Shanmuapriyara, "Heart disease prediction using data mining classification" in *International journal for research in applied science and engineering technology(IJRASET)*, vol.4(2), February 2016

[13] J.Steffi, Dr. R. Balasubramanian and Mr.K.Aravind kumar, "Predicting diabetes mellitus using data mining techniques" in *International journal of engineering development and research (IJEDR)*,vol.6(2), 2018.

[14] Chaitrali S.Dangare and Sulabha S.Apte, "Improved study of heart disease prediction system using data mining classification techniques" in *International journal of computer applications*, vol.47(10), pp.44-48, 2019.

 DOI : 10.35291/2454-9150.2020.0340 © 2020, IJREAM All Rights Reserved.