IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-6, Special Issue-2, March 2018 E-ISSN: 2347-2693

Survey on Disease Diagnostic using Data Mining Techniques

K. Sivaranjani[1*], A. Nisha Jebaseeli[2]

[1*] Department of IT, Bishop Heber College, Bharathidasan University, Tiruchirappalli, TamilNadu, India
[2]Department of CS, Bharathidasan University Constituent College, Tiruchirappalli, TamilNadu, India

Available online at: www.ijcseonline.org

Abstract – Data mining is a large collection of data into knowledge. It is a process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. In data mining, classification is an important function that assigns items in a collection to target categories or classes. The goal of the classification is to accurately predict the target class for each data points. It is a very important technique where large data are classified to retrieve relevant information. There are several classification techniques are available, which includes decision tree algorithm, Bayesian networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithm, adaboost, random forest algorithm and fuzzy logic techniques. This paper proposes the survey of various classification techniques in data mining for healthcare. It also compares the classification techniques and produces the result based on the accuracy level.

Keywords – Data mining, Classification, Decision tree, Bayesian networks, Genetic algorithm.

I. INTRODUCTION

Data mining is a process of knowledge discovery. It is an essential process where intelligent methods are applied to extract the patterns. It is interacting with the user or a knowledge base. Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts. The model is derived based on the analysis of a set of training data (i.e., data objects for which the class labels are known). The model is used to predict the class label of objects for which the class label is unknown. It is a supervised learning.

The functionalities of data mining are used to specify different patterns to mining the tasks. These tasks can be classified into two categories: descriptive and predictive. Tasks are characterizing the properties of the data in a target data set by descriptive mining. Predictive mining tasks perform induction on the current data in order to make predictions. It is used to find out the inference or predict the outcome. Prescriptive data mining uses a combination of techniques and tools to optimize the data that helps to achieve the best outcomes.

Data mining applications have rapidly spread in large sections of organizations offering heath care, in financial predictions, and in weather forecasting. There is a high potential for data mining applications in healthcare. These applications can be grouped, in general, into evaluation of the effects of treatments, management of healthcare, management of relations with patients, and recognition of fraud and misuse. It has an aim to solve real-world problems in diagnosis of diseases, treatments and also in healthcare. The purpose of this survey is to find how different classification techniques are applied in healthcare problems in order to find the accuracy level.

II. RELATED WORK

Many researchers have worked on different classification algorithms for diagnosing the diseases. In this survey paper, diseases diagnosed by various classification techniques such as heart, diabetes, liver and asthma. The following diagram represents various disease diagnoses by classification techniques.

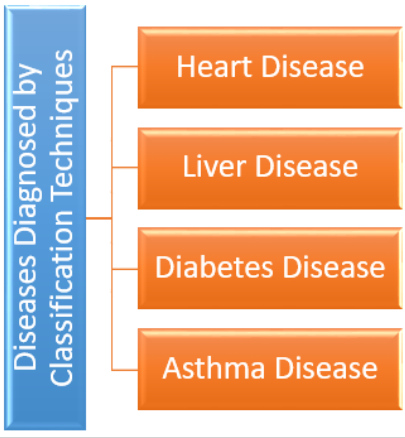

Figure 1. Diseases Diagnosed by Classification Techniques

2.1 Heart Disease

Otoom et al. [10] used the combination of machine learning algorithms. The algorithms proposed coronary artery disease detection and monitoring system. The data set were taken from UCI repository. It consists of 303 rows with 76 attributes. But only 13 attributes are used. For detection of disease there are three algorithms such as Bayes Net, Support Vector Machine and Functional Trees are used. WEKA tool is used for detection. By applying the test on seven best selected attributes, the algorithms Bayes Net, SVM and Function Trees were produced 84.50%, 85.10% and 84.50% of accuracy respectively.

Chaurasia et al. [3] presented the data mining approaches in heart disease detection. WEKA tool was used for mining purposes. Naïve Bayes, J48 and bagging algorithms were used. Data sets were taken from UCI machine learning laboratory. The repository consists of 76 attributes. But only 11 attributes are used for prediction. Naïve Bayes, J48 and Bagging techniques provides 82.31%, 84.35% and 85.03% of accuracy respectively.

There was a hybrid technique proposed by Tan et al. [15], in which genetic and support vector machine algorithms were joined effectively by using wrapper approach. Data sets were collected from UC Irvine Machine learning laboratory for experiment. 84.07% accuracy was obtained by this technique. The graphical representation of accuracy level for this disease is shown in Figure 2.

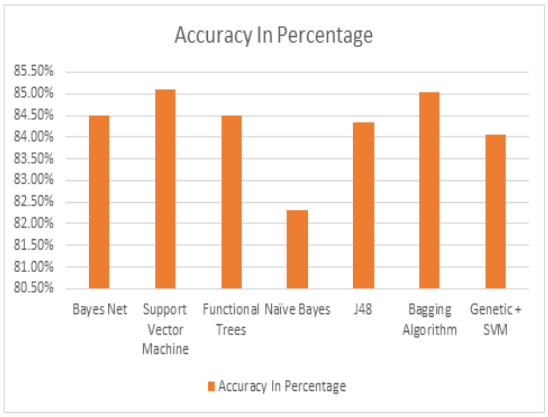

Figure 2. Machine Learning algorithm's accuracy to detect heart disease

2.2 Liver Disease

In [6], Gulia et al. classify the liver patients using five intelligent techniques such as J48, MLP, Random Forest, SVM and Bayesian network. The data was taken from UCI repository. It also involves feature selection method on whole data set to get the correctness of data. After applying Feature Selection method, J48 algorithm provides 70.67% accuracy, MLP algorithm provides 70.84% of accuracy, 71.35% is achieved by SVM technique, Random forest provides 71.87% and Bayes Net given 69.12% of accuracy.

The approach derived by Rajeswari et al. [11] comprises the data mining algorithms called Naïve Bayes, K star and FT tree. Data set consists of 345 rows and 7 features from UCI laboratory. Validation test are applied by using WEKA tool. Naïve Bayes, K star and FT tree produced 96.52%, 83.47% and 97.10% respectively.

In paper [16], Vijayarani et al. predict the disease by SVM and Naïve Bayes classification algorithm. Data set were taken from UCI repository with 560 instances and 10 attributes. These algorithms are also compared based on accuracy and time of execution. 61.28% was obtained from Naïve Bayes with 1670.00ms execution time. SVM obtained 79.66% accuracy with execution time 3210.00ms. MATLAB was used for the implementation part. The following figure shows the graphical view of machine learning algorithms for liver disease detection.

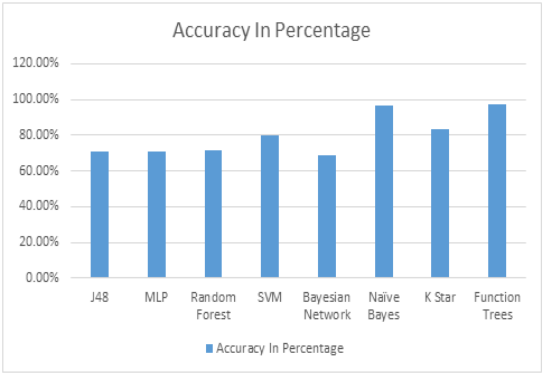

Figure 3. Accuracy of Machine learning algorithms to detect liver disease

2.3 Diabetes Disease

Sarwar et al. [12] suggested the work on Naïve Bayes to predict the diabetes type-2. Type-2 diabetes comes from the growth of Insulin resistance. There are 415 cases are tested with MATLAB and achieved 95% of accuracy.

Iyer et al. [7] performed a work with Decision tree and Naïve Bayes algorithm. Data set taken from Pima Indian Diabetes. WEKA tool is used for test cases. Decision Tree shows 74.87% of accuracy and Naïve Bayes produced 79.56% of accuracy.

Machine learning technique is used by the authors Kumari and Chitra [8]. SVM technique is used with RBF kernel for classifying data. The data set taken from Pima

Indian Diabetes and MATLAB was used for implementation. Accuracy achieved by this SVM is 78%.

Author Ephzibah [5] predict the model for diabetes diagnosis. This model joins the Genetic algorithm and fuzzy logic. UCI laboratory provides data set with 769 instances and 8 attributes. 87% of accuracy obtained from this joined model. The following diagram shows the accuracy graph of algorithms for diabetes disease.

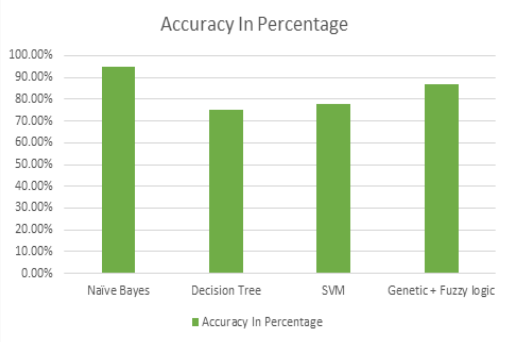

Figure 4. Accuracy of Machine learning algorithms to detect diabetes disease

2.4 Asthma Disease

Nahit Emanet et al. [9] provide the combination of machine learning algorithms such as Random forest algorithm, Adaboost algorithm and Multi-Layer Perceptron algorithm. Asthma prediction using only the chest sound signals using ordinary microphones. Random forest and Adaboost algorithm both provides 90% of accuracy and MLP gets 80% of accuracy.

BDCN Prasad et al. [1], predict Context Sensitive auto-associative memory neural network model (CSAMM), Backpropagation model, C4.5 algorithms, Bayesian network, Particle Swarm Optimization (PSO) techniques. These algorithms were produced 84.32%, 82.21%, 83.83%, 81.17% and 84.16% of accuracy from CSAMM, Backpropagation model, C4.5 algorithms, Bayesian network and PSO techniques respectively.

Taha et al. [14] proposed a method based on K-nearest neighbor, Random forest and SVM algorithms. These are the popular methods for multi-class diagnosis in the area of pattern recognition. 99.34%, 97.37% and 98.70 of accuracy provided by SVM, Random forest and K-nearest neighbor algorithms respectively.

E. Chatzimichail et al. [4] proposed Artificial Neural Network (ANN) approach, MLP algorithm and Probabilistic Neural Networks (PNN) techniques. Prediction algorithm involves two stages: Feature reduction through partial least square (PLS) regression and classification through MLP. All the three techniques produced 96.77% of accuracy. The following diagram presents the graphical representation of asthma prediction disease.

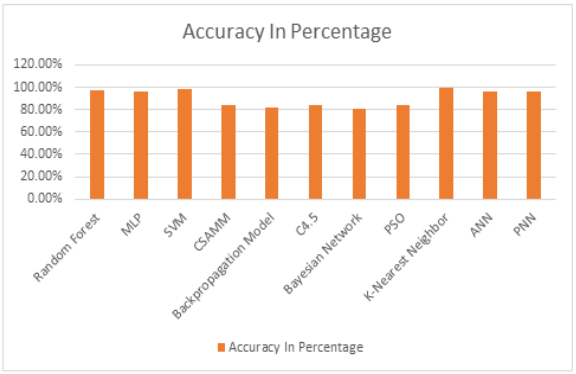

Figure 5. Machine learning algorithms to detect asthma disease

III. PERFORMANCE ANALYSIS

From the analysis of different machine learning techniques for heart, liver, diabetes and asthma disease diagnosis, several algorithms perform very well. From the existing literature, Support Vector Machine and Naïve Bayes techniques provides better accuracy when compared with other algorithms. Artificial Neural Network also very useful in prediction analysis. Tree algorithm is also used to produce enhanced accuracy when it is responded correctly with the attributes of data set.

IV. CONCLUSION

Data mining is used to discover patterns and relationships in the data in order to help and make better decisions. For effective utilization of data mining in health organizations provides to enhance and secure health data sharing. It also provides effective data mining techniques for analyzing data to uncover hidden information. There are several data mining techniques are used in various applications. The statistical models are not capable to produce good performance results for the analysis. Statistical models are not capable to hold categorical data and missing values. It can't focus large data points. But, machine learning plays an important role in many applications such as image detection, natural language processing and disease diagnostics. This paper provides the survey of different machine learning techniques for diagnosing different diseases like heart, liver, diabetes and asthma disease. Many of the algorithms show good results because of identifying the attribute accurately. From this survey, it is observed that the detection of heart disease provides 85.10% of accuracy using Support Vector Machine

algorithm. For diagnose the liver disease, 97.10% of accuracy is achieved using Function Tree algorithm. For detecting diabetes disease, Naïve Bayes algorithm produces 95% of accuracy and K-Nearest Neighbor technique provides 99.34% of accuracy for detecting asthma disease. This survey clearly stated that, the above observed algorithms provide enhanced accuracy on different diseases. Predictive data mining in healthcare is at the forefront of improving quality of care, reducing costs, and improving population health. It has greater potential to drive future models of care and is a key step towards personalized medicine. It also provides opportunity for the improved decision making process.

REFERENCES

[1] BDCN Prasad, P. E. S. N Krishna Prasad and Y Sagar, "An approach to develop expert systems in medical diagnosis using machine learning algorithms (Asthma) and a performance study", International Journal on Soft Computing, Vol. 2, No. 1, pp 26-33, 2011.

[2] J. Cathrin Princy, K.Sivaranjani, "Survey on Asthma Prediction Using Classification Technique", International Journal of Computer Science and Mobile Computing, ISSN: 2320 088X, Vol. 5, No. 7, pp 515 – 518, July 2017.

[3] Chaurasia, V. and Pal, S., "Data Mining Approach to Detect Heart Disease", International Journal of Advanced Computer Science and Information Technology (IJACSIT), Vol. 2, pp 56-66, 2013.

[4] E. Chatzimichail, E. Parakakis and A. Rigas, "Predicting asthma outcome using partial least square regression and artificial neural networks", Advances in Artificial Intelligence, Article ID 435321, pp 1-7, 2013.

[5] Ephzibah, E.P., " Cost Effective Approach on Feature Selection using Genetic Algorithms and Fuzzy Logic for Diabetes Diagnosis", International Journal on Soft Computing (IJSC), Vol.2, pp 1-10, 2011.

[6] Gulia, A., Vohra, R. and Rani, P., "Liver Patient Classification Using Intelligent Techniques", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5, pp 5110-5115, 2014.

[7] Iyer, A., Jeyalatha, S. and Sumbaly, R., "Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol. 5, pp 1-14, 2015.

[8] Kumari, V.A. and Chitra, R., "Classification of Diabetes Disease Using Support Vector Machine", International Journal of Engineering Research and Applications (IJERA), Vol.3, pp 1797-1801, 2013.

[9] Nahit Emanet, Halil R Oz, Nazan Bayram and Dursun Delen, "A comparative analysis of machine learning methods for classification type decision problems in healthcare", Decision Analytics, Vol. 1, No. 6, pp 1-20, 2014.

[10] Otoom, A.F., Abdallah, E.E., Kilani, Y., Kefaye, A. and Ashour, M., "Effective Diagnosis and Monitoring of Heart Disease", International Journal of Software Engineering and Its Applications. Vol. 9, pp 143-156, 2015.

[11] Rajeswari, P. and Reena, G.S., "Analysis of Liver Disorder Using Data Mining Algorithm", Global Journal of Computer Science and Technology, Vol.10, pp 48-52, 2010.

[12] Sarwar, A. and Sharma, V., "Intelligent Naïve Bayes Approach to Diagnose Diabetes Type-2", Special Issue of International Journal of Computer Applications (0975-8887) on Issues and Challenges in Networking, Intelligence and Computing Technologies(ICNICT), Vol.3, pp 14-16, 2012.

[13] A.K. Shafreen Banu, S. Hariganesh, "A Novel Feature Selection Algorithm for Dimensionality Reduction in Microarray Datasets", International Journal of ChemTech Research, ISSN: 2455 9555, Vol. 10, No.14, pp 190-197, 2017.

[14] Taha Samad Soltani Heris, Mostafa Langarizadeh, Zahra Mahmood and, Maryam Zolnoori, "Intelligent diagnosis of Asthma using machine learning algorithms", International Research Journal of Applied and Basic Sciences, Vol. 5, No. 1, pp 140 – 145, 2013.

[15] Tan, K.C., Teoh, E.J., Yu, Q. and Goh, K.C., "A Hybrid Evolutionary Algorithm for Attribute Selection in Data Mining", Journal of Expert System with Applications, Vol. 36, pp 8616-8630, 2009.

[16] Vijayarani, S. and Dhayanand, S., "Liver Disease Prediction using SVM and Naïve Bayes Algorithms", International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, pp 816-820, 2015.

Authors Profile

K.Sivaranjani received her M.Sc., degree in Seethalakshmi Ramaswamy College, Trichy, India in 2002. She also received her M.Phil degree in Bharathidasan University, Trichy, India in 2005. She is cleared SLET in 2016 and NET in 2018. Now she is working as a Assistant Professor with Department of Information Technology, Bishop Heber College, Trichy, India. She is pursuing Ph.D (Computer Science) in Bharathidasan University.

Dr. A. Nisha Jebaseeli obtained her M. Tech degree in Bharathidasan University, Trichy, India in 2008. She obtained her Ph.D (Computer Science) in Bharathidasan University in 2014. Now she is working as a Assistant Professor & head with Department of Computer Science Bharathidasan University Constituent College, Lalgudi, Trichy.

