A SYSTEMATIC SURVEY ON THE APPLICATION OF DIFFERENT DATA MINING AND MACHINE LEARNING ALGORITHMS IN HEALTH CARE SECTOR FOR DISEASE DIAGNOSIS

Asha Unnikrishnan
Asst. Professor, Department of Computer Science,
St. Joseph's College, Devagiri, Calicut, Kerala, India
asha.nkk@gmail.com

Abstract:

Human healthcare became one of the most important topics in the current society. The accurate, fast , effective and robust disease detection is often a difficult task, because of this it became necessary that medicine field searches support from other fields such as computer science and statistics. These disciplines now a days facing the challenge of exploring new techniques, beyond the traditional ones. As large number of techniques emerges every day, it makes necessary to provide a comprehensive overview which avoids the very particular aspects. On this , I propose a systematic review which deals with the application of different Data Mining and Machine Learning algorithms in the diagnosis of human diseases. This review focuses on various modern techniques which are related to the Machine Learning techniques applied to diagnosis of human diseases in the health care field, in order to discover interesting patterns and to make non-trivial predictions in decision-making. I hope, this work can help researchers to discover and determine the applicability of the Machine Learning techniques in their particular specialties. I provide some examples of the algorithms used, analyzes some trends that are focused on the goal searched, and the different area of applications. The advantages and disadvantages of each technique are given to help choose the most appropriate in each real-life situation.

Keywords: Machine Learning; Data Mining; Decision Tree Algorithm; Support Vector Machine; Neural Network.

I. INTRODUCTION

Nowadays, advances of technologies in biological and medical field have been providing us explosive volumes of physiological and biological data. Learning from these data, facilitates the understanding of human health and disease. Particularly, Machine Learning (ML) helps to perform predictive analysis or pattern recognition on large amount of data. Also, it offers a wide range of alerting, risk management and decision support tools, targeted at improving patient's safety and healthcare quality. Machine learning offers a wide range of tools, techniques, and frameworks to address these challenges. There are many algorithms available in ML for data modeling. Some of the important and extensively used machine learning technique for medical diagnosis are Decision Tree Algorithm since its implementation is simple and easy to interpret [1,8]. Some of the other more popular Machine Learning based Algorithm applied for medical diagnosis are Support Vector Machine as well as Artificial Neural Network based techniques [2,14].

II. BACKGROUND OF THE STUDY

Mining of data is the most important stage of information discovery in a database system contains huge amount of data. Data Mining is the process of extracting unique, implicit and potentially useful information from the huge data. The difference between knowledge discovery and Data mining is that the utilization of different intelligent algorithms to excerpt patterns from the data whereas information discovery is the complete process which is involved in discovering knowledge from data. The primary objective is to abstract high-level information from the low-level data . The key approach in research to utilize large volume of data is applying Machine Learning and Data Mining methods for the extraction of knowledge.

III. METHODS

Based on the methodology of a systematic review, I searched in literature databases such as Scopus, Journal Citation Reports (JCR), Google Scholar, and MedLine from the last decade up to the present.

Data Extraction

A protocol for data extraction was defined and evaluated. The following inclusion criteria were used:
Studies with key words such as Human Disease, Heart Disease, Hepatitis , Lever Disease; Diabetes; Dengue belong to the thematic areas Data Mining, Artificial Intelligence, Machine Learning ,Deep Learning, Big Data .The document types are Indexed Journal Papers, Books, Book Chapters, and Conference Papers.
Moreover, the following exclusion criteria were used:
Not among the years 2010-2020; Not belonging to the sub areas of Data Mining, Machine Learning ,Artificial Intelligence, Deep Learning and Big Data ; Not Indexed Journal Paper, Book, Book Chapter, Conference Paper as source type ; Topics addressed in other studies; Not conclusive; Not fully solved.

IV. DIAGNOSIS OF DISEASES BY USING DIFFERENT MACHINE LEARNING ALGORITHMS

Many researchers have worked on various Machine Learning algorithms for diagnosis of diseases. Researchers have been accepted that ML algorithms

performed well in the diagnosis of different diseases. Figurative approach of diseases diagnosed by various Machine Learning Techniques are shown in this paper. This survey paper includes disease diagnosis by ML for Diabetes, Liver disease , Heart disease, Dengue and Hepatitis.

A. HEART DISEASE

Otoom et al . [4] has proposed a system for monitoring and analysis of Coronary artery diseases. Cleveland heart data set is taken from UCI, which consists of 303 cases and 76 attributes/features. Out of 76 features ,13 were used. The algorithm used were Bayes Net, Support Vector Machine (SVM), and Functional Trees (FT) . For detection ,WEKA tool is used. 88.3% accuracy is attained after experimenting Holdout test, by using SVM technique. In Cross Validation test, Both SVM and Bayes Net showed the accuracy of 83.8%. After using FT, 81.5% accuracy was attained . Seven best features are picked up by using Best First selection algorithm. Cross Validation test were used for validation. By applying the test on Seven best selected features, Bayes Net has attained 84.5% of correctness. Support Vector Machine showed 85.1% accuracy and FT classified 84.5% correctly.

Vembandasamy et al . [5] proposed a system, to diagnose heart disease by using Naive Bayes. Naïve Bayes have powerful independence assumption. The data-set used have obtained from one of the leading diabetic research institute in Chennai, which consists of 500 patient's data . WEKA tool was used and executed classification by using 70% of Percentage Split. Naive Bayes offers 86.419% of accuracy.

Chaurasia and Pal [6] proposed a system for heart disease detection. WEKA tool is used. Naive Bayes, J48 and bagging were used . Heart disease data set from UCI machine learning laboratory is used which consists of 76 attributes. Only 11 attributes were employed for prediction purpose. Naive bayes provides 82.31% accuracy. J48 gave 84.35% of correctness and 85.03% of accuracy has achieved by Bagging. On this data set, Bagging offered a better classification rate.

Parthiban and Srivatsa [7] have put their effort in the diagnosis of heart disease in diabetic patients by using the ML Algorithms. Algorithms of Naïve Bayes and SVM are applied by using WEKA tool. Data set consists of 500 patients was used which are collected from Research Institute , Chennai. Out of 500, 142 patients had the disease and 358 did not have the disease. By using Naive Bayes Algorithm 74% of accuracy is attained. In the existing literature Support Vector Machine attained the highest accuracy of 94.60 %. The following Table I shows level of accuracy of different Machine Learning algorithms in the prediction of Heart disease

Table I. ML Algorithms and its accuracy level in the prediction of Heart disease

Algorithm	Accuracy in %
Bayes Net	84.5
SVM	94.6
FT	81.5
Naive Bayes	86.41
J48	84.3
Bagging	85.03

B. DIABETES DISEASE

Iyer et al . [9] has performed a work to predict diabetes disease by using DT(Decision Tree) and Naive Bayes. Pima Indian diabetes data set are the data set used in this work WEKA tool is used to perform tests. In this data-set percentage split of (70:30) predicted better than cross validation. J48 showed 74.8698% and 76.9565% accuracy level by using Cross Validation and Percentage Split Respectively. Naive Bayes attained correctness of 79.5652% using PS.

Sen and Dash [10] proposed a model for the diagnosis of Diabetes disease. Pima Indians diabetes data set which is received from UCI Machine Learning laboratory is used. WEKA tool is used for analysis. CART, Adaboost, Logiboost and grading learning algorithms were also used for the prediction of diabetes . Experimental results were compared based on the correct or incorrect classification. CART offered 78.646% accuracy. The Adaboost offered 77.864% accuracy. Logiboost offered 77.479% accuracy. Grading has the classification rate of 66.406%. CART offered highest accuracy of 78.646% and misclassification Rate of 21.354%, which is lesser as compared to the other techniques.

Kumari and Chitra [11] has also done an experimental work to predict diabetes disease .In their experiment, they used SVM. For classification purpose, RBF kernel is used in SVM. Data set used was Pima Indian diabetes data set provided by Machine Learning laboratory at University of California, Irvine. MATLAB 2010a were used to conduct experiments. SVM offered 78% accuracy level.

Sarwar and Sharma [12] have suggested Naive Bayes for the prediction of Type-2 diabetes .The employed data set consisted of 415 cases and data are gathered from dissimilar sectors of various societies in India. MATLAB with SQL server is used for the development of model. 95% correct prediction has achieved by Naive Bayes.

Ephzibah [13] has also constructed a model for the diagnosis of diabetes. Their proposed model combined GA and fuzzy logic. It is used for the selection of best subset of features and also to enhance accuracy of classification . The employed dataset has picked up from UCI Machine Learning laboratory that has 8 attributes and 769 cases. For the implementation , MATLAB is used. Only three best features/attributes were selected by using Genetic Algorithm. These attributes were used by fuzzy logic classifier which has attained the accuracy level of 87% .It is found that Naive Bayes based systems are helpful for diagnosis of Diabetes disease. It offered a highest accuracy of 95% .The results showed that this system can do good prediction with minimum error to diagnose diabetes

disease. The following Table II shows level of accuracy of different Machine Learning algorithms in the prediction of Diabetes disease

Table II. ML Algorithms and its accuracy level in the prediction of Diabetes disease

Algorithm	Accuracy in %
CART	78.6
Adaboost	94.6
Naive Bayes	95
J48	76.95
Logiboost	77.47

C. LIVER DISEASE

Vijayarani and Dhayanand [15] proposed system to predict the liver disease by using Support vector machine and Naive bayes Classification algorithms. ILPD data set is used which is obtained from UCI. The above mentioned data set comprises of 560 instances and 10 attributes. Comparisons has made on the basis of accuracy level and time of execution. Naive bayes showed 61.28% correctness in 1670.00 ms. SVM attained 79.66% accuracy in 3210.00 ms . MATLAB is used for implementation. SVM showed better accuracy as compared to the Naive bayes .But in the terms of time of execution, Naive Bayes took less time as compared to SVM.

Gulia et al . [16] performed a study on intelligent techniques to classify the liver patients. The data set had chosen from UCI. WEKA tool and five techniques J48, Random Forest, MLP, Bayesian Network and SVM were also used in this experiment. At first step, all algorithms are applied on the original data set and obtained the percentage of correctness. In the second step, Feature Selection method is applied on whole data-set to get the significant subset of liver patients and all the above mentioned algorithms were used to test the subset of the whole data-set. In the third step they took the comparison of outcomes before and after Feature Selection. After FS, algorithms provided accuracy level as J48 had 70.669% accuracy, 70.8405% accuracy has achieved by the MLP algorithm, SVM showed 71.3551% accuracy, 71.8696% accuracy has offered by Random forest and Bayes Net showed 69.1252% accuracy.

Rajeswari and Reena [17] used the algorithms Naive Bayes, K star and FT tree to analyze the liver disease. Data set is taken from UCI which consists of 345 instances and 7 attributes. 10 cross validation test have done by using WEKA tool. Here Naive Bayes provided 96.52% Correctness in 0 sec. 97.10% accuracy has achieved by using FT tree in 0.2 sec. K star algorithm classified the instances with 83.47% accuracy in 0 sec. On the basis of results, highest classification accuracy has offered by FT tree on the liver disease dataset as compared to other algorithms used. Time taken for getting result when algorithm is applied on the dataset of liver disease is low as compared to other algorithms. Which showed the improved performance of FT tree. The following Table III shows level of accuracy of different Machine Learning algorithms in the prediction of Liver disease

Table III. ML Algorithms and its accuracy level in the prediction of Liver disease

Algorithm	Accuracy in %
SVM	79.6
FT	97.1
Naive Bayes	96.5
J48	70.6
Bayes Net	69.12

D. DENGUE DISEASE

Tarmizi et al . [22] have performed a work on the detection on Malaysia Dengue outbreak .The highly contagious disease like Dengue creates trouble mostly in the countries like Thailand, Indonesia and Malaysia. Decision Tree (DT), Rough Set Theory (RS) and Artificial Neural Network(ANN) were the classification algorithms used in their study . Data set are taken from Public Health Department of Selangor State. WEKA data mining tool has used. 10Cross-fold Validation and Percentage split test has done. In 10-Cross fold validation , Decision Tree offered the accuracy of 99.95% , Artificial Neural Network attained 99.98% of accuracy and RS showed 100% of accuracy level. After using PS, Both Decision tree (DT) and Artificial Neural Network (ANN) provided 99.92% of correctness. And RS achieved 99.72% accuracy level.

Fathima and Manimeglai [18] also performed a work on the prediction of Arbovirus-Dengue disease. The algorithm used by these researchers in this work was Support Vector Machine(SVM). Data set has obtained from King Institute of Preventive Medicine and from various hospitals and laboratories of Chennai and Tirunelveli from India. It consists of 29 attributes and 5000 samples. Data has examined by R project version 2.12.2. The obtained accuracy level by SVM was 90.42 %. The following Table 4 shows level of accuracy of different Machine Learning algorithms in the prediction of Dengue disease

Table IV. ML Algorithms and its accuracy level in the prediction of Dengue disease

Algorithm	Accuracy in %
SVM	90.4
DT	99.9
RS	100

E. HEPATITIS DISEASE

Ba-Alwi and Hintaya [19] has done a comparative analysis on various data mining algorithms like Naive Bayes, K Star, FT Tree, LMT, J48, and NN for the diagnosis of Hepatitis Disease. The data set was taken

from UCI Machine Learning repository. Results are measured and classified in terms of accuracy and time. Comparative Analysis is taken by using WEKA and neural connections tool. In this Analysis they used a second technique Rough Set theory, by using WEKA. Performance of RS was higher than NN .The accuracy level of various algorithsm are Naive Bayes gave 96.52% in 0sec. 84% was attained by the Naive Bayes Updateable algorithm in 0 sec. FT Tree attained 87.10%. in 0.2 sec and K star offered 83.47% of correctness in 0 sec. 83% has achieved by J48 in 0.03 sec. LMT provides 83.6% in 0.6 sec. Neural network showed 70.41% .

Karlik [20] gave a comparative analysis of Naive Bayes and back propagation for the diagnosis of hepatitis disease. The different types of hepatitis "A, B, C, D and E" are generated by different viruses . the have used RapidMiner open source software in their analysis. The Hepatitis data set has taken from UCI ,which include 20 features and 155 instances. In this experiment, 15 attributes were used. Naive Bayes classifier gave 97% accuracy. Three-layered feed forward Neural Network were used and which was trained with Back propagation algorithm .For training purpose, 155 instances were used . Feed forward Neural Network with back propagation showed highest accuracy of 98%. The following Table 5 shows level of accuracy of different Machine Learning algorithms in the prediction of Hepatitis disease

Table V. ML Algorithms and its accuracy level in the prediction of Hepatitis disease

Algorithm	Accuracy in %
Naïve Bayes	96.5
FT	87.1
LMT	83.6
NN	70.4
FFNN	98

V. DISCUSSIONS AND ANALYSIS OF MACHINE LEARNING TECHNIQUES

For diagnosis of Heart disease, Diabetes, Liver disease, Dengue and Hepatitis , various ML algorithms performed well. From existing literature, it is observed that Naive Bayes and SVM algorithms were widely used in the detection of various diseases. Both algorithms offered better accuracy level as compared to other algorithms. Artificial Neural network (ANN) is also very useful for the prediction of diseases. ANN showed good outcome but it took more time as compared to other algorithms. FT Tree algorithm were also used but they did not attain wide acceptance due to its complexity. It is also found RS theory is not widely used but it presents maximum output. The following Figure 1 shows a comprehensive graph of highest accuracy level attained by different Machine Learning algorithms for the prediction of Heart disease, Diabetes , Liver Disease ,Dengue and Hepatitis

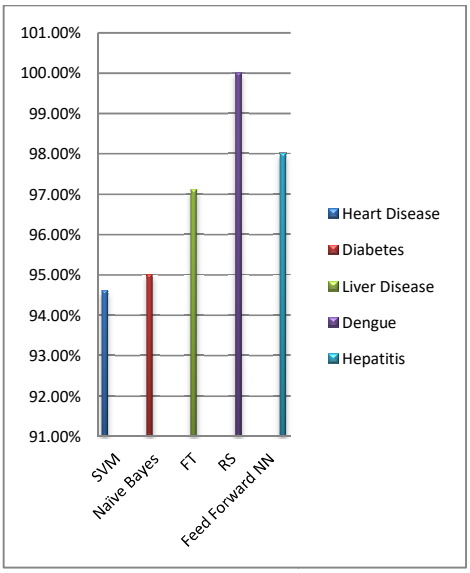

Fig 1. Accuracy level attained by different ML algorithms for the prediction of different diseases

VI. CONCLUSION

Machine Learning plays a vital role in various applications like Image detection, Data mining, Natural Language Processing, and Disease Diagnostics. Machine Learning offers possible solutions in these different domains. This paper provides a survey on different Machine Learning techniques for the diagnosis of different Diseases like Diabetes ,Heart disease, Liver disease, Hepatitis and Dengue .Many algorithms have shown good performance and results since they identify the attribute accurately. From the previous study, it is observed that for the detection of Heart disease, SVM provides accuracy level of 94.60%. For the detection of Diabetes , by Naive Bayes offered the highest classification accuracy of 95%. For the diagnosis of Liver disease ,FT provides 97.10% of correctness . 100% accuracy is achieved by RS theory in dengue disease detection. The Feed Forward Neural Network correctly classifies hepatitis disease with 98% of accuracy level. Figure 1 shows comprehensive graph of highest accuracy level attained by different Machine Learning algorithms for the prediction of the above mentioned diseases . From analysis, it is clearly observed that these algorithms provide enhanced accuracy on the prediction of different diseases. This paper provides a set of tools which are developed in the community of AI. These tools are really useful for the analysis of problems and also provide opportunity for the improved decision making process.

REFERENCES

[1] Iniesta, R., Stahl, D., & McGuffin, P. (2016). Machine learning, Statistical learning and the future of biological research in psychiatry. *Psychological medicine*, *46*(12), 2455- 2465.

[2] Karamizadeh, S., Abdullah, S. M., Halimi, M., Shayan, J., & javad Rajabi, M. (2014, September). Advantage and drawback of support vector machine functionality. In *2014 international conference on computer, communications, and control technology (I4CT)* (pp. 63- 65). IEEE.

[3] Babu, S., Vivek, E. M., Famina, K. P., Fida, K., Aswathi, P., Shanid, M., & Hena, M. (2017, April). Heart disease diagnosis using data mining technique. In *2017 international conference of electronics, communication and aerospace technology (ICECA)* (Vol. 1, pp. 750-753). IEEE.

[4] Otoom, A. F., Abdallah, E. E., Kilani, Y., Kefaye, A., & Ashour, M. (2015). Effective diagnosis and monitoring of heart disease. *International Journal of Software Engineering and Its Applications*, *9*(1), 143-156.

[5] Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart diseases detection using Naive Bayes algorithm. *International Journal of Innovative Science, Engineering & Technology*, *2*(9), 441-444.

[6] Chaurasia, V., & Pal, S. (2014). Data mining approach to detect heart diseases. *International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol*, *2*, 56-66.

[7] Parthiban, G., & Srivatsa, S. K. (2012). Applying machine learning methods in diagnosing heart disease for diabetic patients. *International Journal of Applied Information Systems (IJAIS)*, *3*(7), 25-30.

[8] Tan, K. C., Teoh, E. J., Yu, Q., & Goh, K. C. (2009). A hybrid evolutionary algorithm for attribute selection in data mining. *Expert Systems with Applications*, *36*(4), 8616-8630.

[9] Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *arXiv preprint arXiv:1502.03774*.

[10] Sen, S. K., & Dash, S. (2014). Application of meta learning algorithms for the prediction of diabetes disease. *International Journal of Advance Research in Computer Science and Management Studies*, *2*(12).

[11] Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications*, *3*(2), 1797-1801.

[12] Sarvwar, A., & Sharma, V. (2012). Intelligent Naive Bayes approach to diagnose diabetes type-2. Special Issue Int. *J. Comput. Appl. Issues Challeng. Netw. Intell. Comput. Technol*.

[13] Ephzibah, E. P. (2011). Cost effective approach on feature selection using genetic algorithms and fuzzy logic for diabetes diagnosis. *arXivreprint arXiv:1103.0087*.

[14] Archana, S., & Elangovan, K. (2014). Survey of classification techniques in data mining. *International Journal of Computer Science and Mobile Applications*, *2*(2), 65-71.

[15] Vijayarani, S., & Dhayanand, S. (2015). Liver disease prediction using SVM and Naïve Bayes algorithms. *International Journal of Science, Engineering and Technology Research (IJSETR)*, *4*(4), 816-820.

[16] Gulia, A., Vohra, R., & Rani, P. (2014). Liver patient classification using intelligent techniques. *International Journal of Computer Science and Information Technologies*, *5*(4), 5110-5115.

[17] Rajeswari, P., & Reena, G. S. (2010). Analysis of liver disorder using data mining algorithm. *Global journal of computer science and technology*.

[18] Fathima, A., & Manimegalai, D. (2012). Predictive analysis for the arbovirus-dengue using svm classification. *International Journal of Engineering and Technology*, *2*(3), 521-7.

[19] Ba-Alwi, F. M., & Hintaya, H. M. (2013). Comparative study for analysis the prognostic in hepatitis data: data mining approach. *International Journal of Scientific & Engineering Research*, *4*(8), 680-685.

[20] Karlik, B. (2012). Hepatitis disease diagnosis using backpropagation and the naive bayes classifiers. *IBU Journal of Science and Technology*, *1*(1).

[21] Sathyadevi, G. (2011). Application of CART algorithm in hepatitis disease diagnosis. In *2011 International Conference on Recent Trends in Information Technology (ICRTIT)* (pp. 1283-1287). IEEE.

[22] Tarmizi, N. D. A., Jamaluddin, F., Abu Bakar, A., Othman, Z. A., Zainudin, S., & Hamdan, A. R. (2013). Malaysia Dengue Outbreak Detection Using Data Mining Models. *Journal of Next Generation Information Technology (JNIT)*, *4*(6), 96-107.