ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

USE OF MACHINE LEARNING TECHNIQUES TO EFFECTIVELY MANAGE AND DIAGNOSE HUGE AMOUNT OF DATA IN THE FIELD OF HEALTH CARE INDUSTRY

Parshva Jain[*1], Rahul Vijayvargiya[2] & Arsheen[3]
[*1&2]Information Technology, Medicaps, Indore (M.P.)-452001, India
[3]Electronics, Acropolis Technical Campus, Indore (M.P.), India

DOI: 10.5281/zenodo.1042068

ABSTRACT
Health Care industries are generating huge amount of data on the daily basis. It becomes cumbersome task to diagnose the disease. Government is also paying attention on the drugs and human health. All health care industries will collect the large amount of medical data from the human existence to improve the understanding so that proper diagnosis for some disease is possible. An appropriate and accurate computer based automated decision support system is required to reduce cost for achieving clinical tests. This paper provides an insight into machine learning techniques used in diagnosing various diseases. Various data mining classifiers have been discussed which has emerged in recent years for efficient and effective disease diagnosis

KEYWORDS: Health Care Industries, Disease, Diagnosis, cost, Machine Learning Techniques.

I. INTRODUCTION
Healthcare has become one of India's largest sectors - both in terms of revenue and employment. Healthcare comprises hospitals, medical devices, clinical trials, outsourcing, telemedicine, medical tourism, health insurance and medical equipment. The Indian healthcare sector is growing at a brisk pace due to its strengthening coverage, services and increasing expenditure by public as well private players. Indian healthcare delivery system is categorised into two major components - public and private [1].The Government, i.e. public healthcare system comprises limited secondary and tertiary care institutions in key cities and focuses on providing basic healthcare facilities in the form of primary healthcare centers (PHCs) in rural areas. The private sector provides majority of secondary, tertiary and quaternary care institutions with a major concentration in metros, tier I and tier II cities. India's competitive advantage lies in its large pool of well-trained medical professionals. India is also cost competitive compared to its peers in Asia and Western countries. The cost of surgery in India is about one-tenth of that in the US or Western Europe.

These days, health care industry generates a large amount of complex data about patients regarding clinical examination, treatment report, hospital resource management records, electronic patient records, medicine etc which has become cumbersome to organize properly. Due to improper organization of the data, the quality of decision making is getting affected [1]. This increase in data volume requires ways in which data can be extracted and processed efficiently. The accurate diagnosis of life-threatening diseases such as breast cancer, heart disease, liver disease etc is a very crucial task in medical science. The humans and computers can be integrated together to achieve best results for correct diagnosis of diseases by balancing the knowledge of human experts in related domains with the vast search potential of computers. This kind of difficulty could be resolved with the help of machine learning techniques. Computer based decision support system can play an important role in correct diagnosis and cost effective treatment.

The use of computers and information technology is being increasingly implemented in health care organization in order to help doctors in their day to day decision making activities. It helps doctors and physicians in diseases management, tests, medications and discovery of patterns and relationships among clinical and diagnosis data and as well as employ machine learning techniques.

This paper is organized as follows: Section II gives the Categorization of Machine Learning techniques including supervised and unsupervised algorithms. Section III describes the work in the literature regarding Categorization algorithms for medical diagnosis of diseases. Section IV concludes the survey.

II. CATEGORIZATION OF MACHINE LEARNING TECHNIQUES

According to Levi Thatcher (Director of Data Science, Health Catalyst) that throughout healthcare, and many other industries, there are heuristics and established best practices that help people make decisions. A popular example in healthcare is the **LACE index**, which provides the likelihood of patient 30-day readmission risk. You might have also heard of similar tools like the **SOFA Score**, **Apgar Score**, **PRISM Score**, and the **PIM Score**.

Like most of these scores, the **LACE calculation** is fairly simple. It's based on length of stay, acuity of the admission, patient comorbities, and ED visits within the last six months. In each of these categories, points are assigned—a length of stay of three days equals three points, for example. Then the points from each categories are added up to form the LACE index.

It's simple and indicative of how healthcare has worked for the last 20-30 years. First, there's a national study, which eventually leads to guidelines and a simple calculation to help prioritize which patients are most at risk of something.

So what's wrong with that? Well, the guidelines can only be **narrowly applied** and even then **don't give impressive results**. Think of it—LACE was developed from patients seen in Ontario from 2004 to 2008. Do your patient demographics closely match those in Ontario? Or, do your patient demographics even match your same set from ten years ago? Perhaps not. Another issue is applicability—since LACE requires the patient's length of stay, the score is only available upon discharge. What if you want a risk score early during their stay?
This is why machine learning is fantastic—it fills these gaps. First, it learns the important relationships in your data on past patients and their outcomes. This means that the model is customized on your data from the last few years–you don't have to rely on scores made on other populations, 10-20 years ago. Second, machine learning allows you to create a model based on whatever data is available when you need a risk score (i.e., upon admission rather than discharge).

Machine learning is a domain of artificial intelligence involving the construction of algorithms that automatically learns through experience and performance of algorithm gets improved with each experience [2]. Algorithm operates by detecting some pattern in input data and building a model based on input data to make precise predictions for new data. The machine learning techniques are based on identifying patterns from large data sets that provide support for predictions and decision making process for diagnosis and treatment planning. The machine learning techniques can be classified as follows:

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

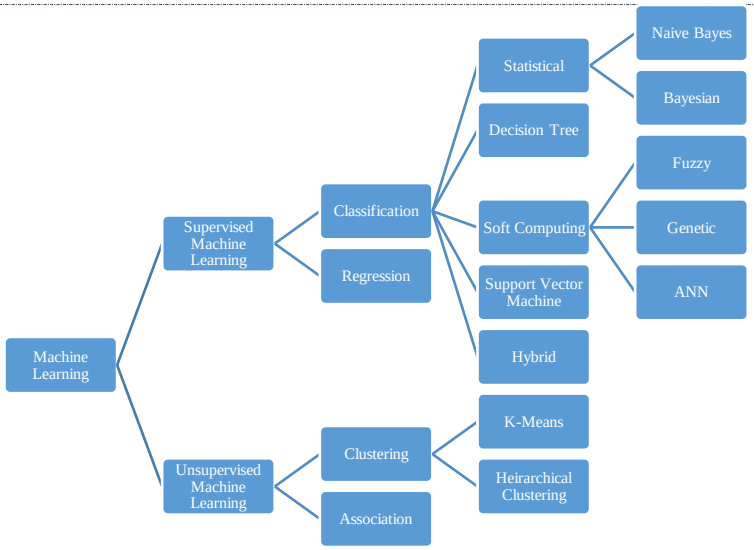

Figure 1: Categorization of Machine Learning Techniques

III. LITERATURE WORK

Work done by various researchers in the field of disease diagnosis using data mining and machine learning techniques has been discussed below. Decision tree has been used as classifier for breast cancer diagnosis [3-5]. Azar and El-Metwally [6] proposed a decision support tool for the detection of breast cancer based on three different classifiers, namely, single decision tree (SDT), boosted decision tree (BDT) and decision tree forest. They claimed that BDT performed better than SDT with 98.83% and 97.07% accuracy respectively. The demerit of decision tree classifier in medical diagnosis is imbalancement and cost sensitivity problem.

1. Chowdhury et al. [7] utilized ANN in predicting neonatal disease diagnosis. The proposed technique comprised of Multi Layer Perceptron with a back propagation learning algorithm for training ANN and recognizing a pattern for the diagnosing and prediction of neonatal diseases. The data set consists of 94 samples of different symptoms parameter. The technique exhibits ANN based prediction of neonatal disease with an accuracy of 75% with 64 training set and, 15 test set and 15 validation test.
2. Vanisree et al. [8], proposed a Decision Support System for diagnosis of Congenital Heart Disease. The proposed system is based on Backpropagation neural network which is multi layered Feed Forward Neural Network, which is trained by a supervised Delta Learning Rule. The dataset consists of 200 samples with 36 attributes each depicting signs, symptoms and the results of physical evaluation of a patient. The proposed system used 80% data set for training and 20% for testing and achieved an accuracy of 90% and mean square error of 0.016.
3. Ratnakar et al. [9] proposed a solution based on genetic algorithm for selection of optimal set of attributes for prediction of heart diseases and Naïve Bayes‟ technique to generate relationships amongst the attributes using the concepts of conditional probability. Using GA, 13 attributes are reduced to 6 which are further fed to Naïve Bayes Classifier.
4. Chitra et al. [10] claimed that the application of ANN can be time-consuming due to the selection of input features for the multi layer perceptron. It is very slow training process and clinicians find it difficult to understand how its classification decisions are taken and cannot interpret the results easily.
5. Masethe [11] performed a comparison of various data mining algorithms on WEKA tool for the prediction of heart attacks to find the best method of prediction. The algorithms used are J48, REPTREE, Naïve Bayes, Bayes net and CART with prediction accuracy as 99.07%, 99.07%, 97.22%, 98.14%, 99.07% respectively.
6. Archana and Sandeep [12] proposed a hybrid prediction model with missing value imputation (HPM-MI) based on K-means clustering with Multilayer Perceptron. The proposed algorithm was evaluated on three medical data sets, Pima Indians Diabetes, Wisconsin Breast Cancer, and Hepatitis from the UCI Repository of Machine Learning. The results show HPM-MI has produced accuracy, sensitivity, specificity, kappa and ROC as 99.82%, 100%, 99.74%, 0.996 and 1.0 respectively for Pima Indian Diabetes data set, 99.39%, 99.31%, 99.54%, 0.986, and 1.0 respectively for breast cancer data set and 99.08%, 100%, 96.55%, 0.978 and 0.99 respectively for Hepatitis data set.

7. Turabieh [13] proposed a hybrid algorithm which integrates two powerful computational intelligence techniques namely Gray Wolf Optimization (GWO) and Artificial Neural Networks (ANN) for prediction of heart disease [17]. Gray wolf optimization is a global search method that works by minimizing the root mean square error while gradient-based back propagation method is a local search one. GWO is used for finding the initial optimal weights and biases for ANN model to reduce the probability of ANN getting stuck at local minima and slowly converging to global optimum. The performance of hybrid ANN-GWO is compared with normal ANN trained using back-propagation neural network. The results shown depict that the proposed model increases the convergence speed (time reduces to half) and the prediction accuracy.
8. Tina Patil et al. [14] have applied two classification algorithm viz. Naïve Bayes based on probability and J48 based on decision trees to classify the item according to its features with respect to the predefined set of classes. The results demonstrate that J48 is more accurate and cost efficient than Naive Bayes algorithm.
9. Sunila et al. [15] proposed an improved Multilayer perceptron algorithm (MLP) which works on multiple subsets of training set. The majority probability rule is used to combine the results from different subsets. The experiment is implemented with 10-fold cross validation on Cleveland, Switzerland and Hungarian datasets. The result shows that proposed approach is better than MLP algorithm and has attained an accuracy of 82.8%.
10. Zheng et al [16] proposed K-means algorithm to recognize the hidden patterns of the benign and malignant tumors separately. The membership of each tumor to these patterns is calculated and treated as a new feature in the training model. Then, a support vector machine (SVM) is used to obtain the new classifier to differentiate the incoming tumors. The proposed algorithm achieves the accuracy to 97.38% with 10-fold cross validation.

IV. CONCLUSION

Clearly, the most important priorities for medical research are development of more effective health delivery strategies for developing countries and control of the common and intractable communicable diseases. In this context, the argument has been that much of the medical research that has been carried out in industrial countries, with its focus on non communicable disease and its outcomes in high-technology practice, is completely irrelevant to the needs of developing countries. This view of the medical scene, however, is short term. Although some redistribution of effort is required, every country that passes through the epidemiological transition is now encountering the major killers of industrial countries. Learning more about those killers' basic causes, prevention, and management is crucial. Although the initial costs of providing the benefits of this research are often extremely high, they tend to fall as particular forms of treatment become more widely applied. Hence, because we cannot completely rely on our current preventive measures to control these diseases, medical research must continue.

This survey provides the brief description of machine learning techniques for classification of diseases. The classification accuracy depends on the exact metrics which are used which also indicates the variety of features has been utilized. The role of classifier is important in healthcare industry so that the results can be used for determining the treatment. The existing techniques are studied and compared for finding the efficient and accurate systems.

V. REFERENCES

[1] Department of Industrial Policy and Promotion (DIPP), RNCOS Reports, Media Reports, Press Information Bureau (PIB), Union Budget 2017-18.
[2] Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. eds., 2013. Machine learning: An artificial intelligence approach. Springer Science & Business Media.
[3] D. M. F. bin Othman and T. M. S. Yau, “Comparison of Different Classification Techniques Using WEKA for Breast Cancer,” in 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, F. Ibrahim, N. A. A. Osman, J. Usman, and N. A. Kadri, Eds. Springer Berlin Heidelberg, 2007, pp. 520–523.
[4] N. Cruz-Ramírez, H.-G. Acosta-Mesa, H. Carrillo-Calvet, and R.-E. Barrientos-Martínez, “Discovering interobserver variability in the cytodiagnosis of breast cancer using decision trees and Bayesian networks,” Appl. Soft Comput., vol. 9, no. 4, pp. 1331–1342, Sep. 2009.
[5] C.-Y. Fan, P.-C. Chang, J.-J. Lin, and J. C. Hsieh, “A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification,” Appl. Soft Comput., vol. 11, no. 1, pp. 632–644, Jan. 2011.

[6] A. T. Azar and S. M. El-Metwally, "Decision tree classifiers for automated medical diagnosis," Neural Comput. Appl., vol. 23, no. 7–8, pp. 2387–2403, Dec. 2013.
[7] D. R. Chowdhury, M. Chatterjee & R. K. Samanta, "An Artificial Neural Network Model for Neonatal Disease Diagnosis", International Journal of Artificial Intelligence and Expert Systems (IJAE), vol. 2, no. 3, pp. 96-106, 2011.
[8] Vanisree K, Jyothi Singaraju, "Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks", International Journal of Computer Applications, vol. 19, no. 6, pp. 6-12, 2011.
[9] S. Ratnakar, K. Rajeswari & R. Jacob, "Prediction of Heart Disease Using Genetic Algorithm for Selection of Optimal Reduced Set of Attributes", International Journal of Advanced Computational Engineering and Networking, vol. 1, no.2, pp. 51-55, 2013.
[10] Anuja Kumari, V & Chitra, R 2013, „Classification of Diabetes Disease Using Support Vector Machine", International Journal of Engineering Research and Applications, vol. 3, no. 2, pp. 1797-1801.
[11] Masethe, H.D., Masethe, M.A.: Prediction of heart disease using classification algorithms. In: World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, San Francisco, USA, 22–24 Oct 2014.
[12] Purwar, A. and Singh, S.K. (2015) Hybrid Prediction Model with Missing Value Imputation for Medical Data. Expert Systems with Applications, 42, 5621-5631.
[13] Turabieh, H. , "A Hybrid ANN-GWO Algorithm for Prediction of Heart Disease", American Journal of Operations Research, 6, pp. 136-146, 2016.
[14] Tina Patil, R & Sherekar, SS 2013, „Performance Analysis of Naive bayes and J48 Classification Algorithm for Data Classification", International Journal of Computer Science and Applications, vol. 6, no.2, pp. 256-261.
[15] P. Panday and N. Godara, "Decision Support System for Cardiovascular Heart Disease Diagnosis using Improved Multilayer Perceptron," International Journal of Computer Applications ,vol. 45, no. 8, pp. 12–20, 2012.
[16] Zheng, B., Yoon, S.W. and Lam, S.S., 2014. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 41(4), pp.1476-1482

CITE AN ARTICLE

Jain, P., Vijayvargiya, R., & Barnagarwala, A. (2017). USE OF MACHINE LEARNING TECHNIQUES TO EFFECTIVELY MANAGE AND DIAGNOSE HUGE AMOUNT OF DATA IN THE FIELD OF HEALTH CARE INDUSTRY. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY*, 6(11), 17-21.