# A STUDY OF MACHINE LEARNING ALGORITHMS FOR DISEASE AND EPIDEMIC PREDICTION

[1]AKANSHA DUA, [2]ARPIT KHANNA

*[1,2] B.Tech Student, Department of Computer Engineering, Bharati Vidyapeeth University College of Engineering Pune, Maharashtra, India*

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract—** *With a tremendous growth of big data in the field of healthcare, there has been a significant development made in predicting diseases with the application of machine learning algorithms. From the prediction of epidemic outbreak and multiple diseases to providing better means of storing and securing healthcare data, implementation of machine learning in the field of healthcare promises accurate results. The main focus is on to use machine learning in healthcare to supplement patient care for better results. Machine learning has made easier to identify different diseases and diagnosis them correctly. Predictive analysis with the help of efficient multiple machine learning algorithms helps to predict the disease more correctly and help treat patients. Collecting medical data of humans using IoT and evaluating and analyzing them to predict the disease has provided an efficient mechanism for an integrated machine learning and IoT solution in healthcare. This paper mainly aims to provide collective mechanisms which would implement machine learning technologies to yield accurate results. This purpose of this paper is to study optimized machine learning algorithms for their utilization in multiple disease prediction.*

***Keywords — epidemic outbreak; diagnosis; healthcare; machine leaning algorithms; predicting diseases.***

## I. INTRODUCTION

Today's computing era is the era of the machine learning. Tremendous and rapid increase in data storage and computer processing has resulted in the development of different, magical and smarter form of technology. Earlier in the tabulating systems era (1900s — 1940s) and programming era (from 1950s), computer systems were only able to perform iterative, conditional and logical computations. However with the advent of new technologies, such computing systems were developed which could learn without being explicitly programmed to perform specific task. Such technology is termed as Machine Learning where a machine or a system is able to learn from certain algorithms and make predictions on the basis of it.

Machine Learning algorithms are implemented everywhere. It has become increasingly popular with more and more applications in places where we cannot even think of. Its implementation can be found in many fields. You would be using it in one of its form and you wouldn't even know. That is the magic of it. From virtual personal assistants, predictions while commuting, video surveillance and social media services to online customer support, search engine result refining, product recommendations and online fraud detection, machine learning algorithms are widely used.

One such implementation of machine learning algorithms is in the field of healthcare. Medical facilities need to be advanced so that better decisions for patient diagnosis and treatment options can be made. Machine learning in healthcare aids the humans to process huge and complex medical datasets and then analyze them into clinical insights. This then can further be used by physicians in providing medical care. Hence machine learning when implemented in healthcare can leads to increased patient satisfaction.

In this project, we try to implement multiple functionalities of machine learning in healthcare in a single system. Instead of diagnosis, when a disease prediction is implemented using certain machine learning predictive algorithms then healthcare can be made smart. Some cases can occur when early diagnosis of a disease is not within reach. Hence disease prediction can be effectively implemented. As widely said "Prevention is better than cure", prediction of diseases and epidemic outbreak would lead to an early prevention of an occurrence of a disease. This project mainly focus on the development of a system or we could say an immediate medical provision which would incorporate the symptoms collected from multisensory devices and other medical data and store them into a healthcare dataset. This dataset would then be analyzed using certain efficient machine learning algorithms to deliver results with maximum accuracy.

## II. RELATED WORK

There have been many studies in the field of implementing machine learning techniques in healthcare. Machine learning in healthcare has been one of the top priorities for the researchers. Using various data mining techniques and studying about the hidden patterns, insights can be extracted. These insights can further be used for disease and epidemic prediction.

Datasets from the already available repositories can be used for training the machine. Data pre-processing or data cleaning is one the important aspect to be carried out before implementing machine learning algorithms for mining purposes. We need to systematize or normalize the data for a better understanding. [1]. Paper [2] describes how weighted fuzzy rule based system can be used for diagnosis of heart disease. This paper also implements the neural network based system. Attribute selection from the dataset is one the relevant aspects of data mining. Paper [3] brings into light of selecting the minimum and suitable features. This helps in increasing the prediction accuracy and decreasing the computational complexity. Research work by paper [6] also describes various attribute selection techniques. It concludes that both Adaboost and Decision tree are highly efficient classification algorithms for diseases prediction and yields good performance. This [5] paper analyzes the resource frequency prediction and states how highly frequented resources yields better results of prediction than lowly frequented resources. It also compares the performance of the Nearest Neighbours and Random Forest.

Paper [6] proposes a model based on neural network. This paper aims to achieve a prediction of 94.8%. When compared with the convergence speed of the CNN- based model, the convergence speed of this model is relatively greater. The paper [7] has summarized various data mining techniques such as classification, clustering, association to analyze and predict human disease. Another noteworthy research by paper [8] has been made which defines a system that can prevent spread of airborne diseases. Using big data technologies and IoT, an intelligent dispenser can be made which could prevent millions of dollars spent on infectious diseases.

Paper [9] describes how Re-RX with J48graft algorithm can be used to predict diabetes with high accuracy. To predict type II diabetes, a comparison has been made in paper [10] among the various data mining classification algorithms like Naive Bayes, RFB Network and J48 who are having the accuracy of 76.95%, 74.35% and 76.52% respectively. Paper [11] defines and compares decision tree, random forest, Naive Bayes and Support Vector machine to predict type II diabetes. It uses fivefold cross validation approach for the comparison.

The objective of the project is to implement multiple machine learning algorithms to ease the patient care and treatment. Machine learning methods can be smartly used to improve patient and health care.A smart future can be build in the field of healthcare by implementing the suitable machine learning algorithms.

## III FEATURE SELECTION

Healthcare data used for the training of the system might consist of a number of attributes. It may also contain noise, not allowed and non-relevant data. Preprocessing of the training dataset needs to be done so that data can be cleaned. This enhances the performance and reduces the time for building the model. Efficiency of the algorithm also depends on the attributes used for mining the dataset. Not all attributes in the dataset may prove to be crucial for the model building. Some irrelevant attributes or features may only add the time overhead. Hence selecting of the features is an important step to be considered in the process of prediction.

The main aim behind the process of feature selection is to remove the redundant and non relevant data. This results in improving the efficiency of classifier and hence increases the accuracy in percentage of true positive predictive values. Non relevant features results in the decrease of accuracy.

## IV MACHINE LEARNING TECHNIQUES

Machine learning is solely the process of developing a system that can learn from its past experience. Machine learning is a mechanism that covers data mining and statistics. If a computer system has the capability to improve its performance of computing or performing a certain task on the basis of its past experience, better decisions can be made on patient's diagnoses and treatment options, by applying suitable machine learning technique. Some of its techniques are explained below.

### A. Decision Tree

A decision tree as defined by Han and Kamber (2006) [12] is a visual chart which is represented in the form of a hierarchical structure. The topmost node in the tree represents the root tree. After applying a splitting algorithm, an internal node is generated which represents a test on an attribute.

Decision tree learning algorithm is a supervised machine learning algorithm. It uses a decision tree which is used to map items to certain predictive conclusions. It is based on the classification model which categorizes the input data into an outcome.

This outcome may belong to a particular class domain. The training medical dataset is used to build a classification model in the form of a decision tree. This model is then further used to map the testing dataset to a predictive result. The accuracy of the algorithm depends on the features selected to train the machine.

Splitting attribute can be selected using Information Gain, Gini Index and Gain Ratio Decision Trees.

- Information Gain selects the attribute in such way that entropy is maximized. It gives the measure of how much information feature can provide about a class.

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

- Gini index provide the measure of impurity of a data.

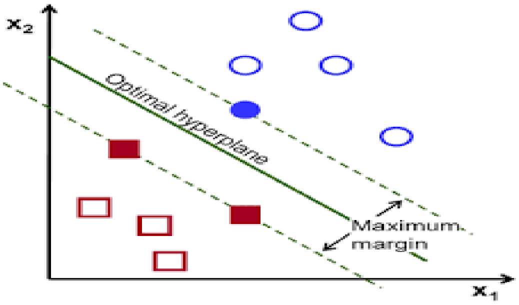
$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

- Gain Ratio is the ratio of information gain (entropy) to the split information.

$$\text{Gain Ratio} = \frac{\text{information gain}}{\text{split information}}$$

## B. Support Vector Machine

Support Vector Machine looks at the extreme of the datasets and draws a decision boundary also known as hyper plane. It is a technique which best segregates the two classes. Both pattern classification and nonlinear regression can be implemented using Support Vector Machine (SVM). The main idea behind the Support Vector Machine is to generate a multidimensional hyper-plane. This hyper-plane can further be used to discriminate between two classes. When the amount of input variables is larger comparative to the available observations, then Support Vector Machine (SVM) can implemented with a greater ease than the other classification algorithms.



As given in paper [13], the hyper-planes can be constructed as follows:

Let in the given training dataset {($x_i$, $y_i$)} where i range from 1 to N, $x_i$ is the input pattern for the $i_{th}$ instance and the target output being represented as $y_i$. We can linearly separate patterns which are represented by the subset $y_i$ = +1 and the subset $y_i$ = -1. The equation of the hyper-plane of this separation is represented as

$$w^T x + b = 0$$

Here x represents input vector, w represents an adjustable weight factor and b is a bias. Thus,

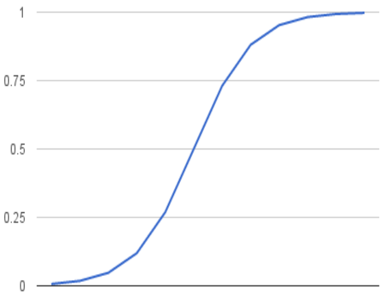
$$w^T x + b >= 0 \quad \text{for} \quad y_i = +1$$
$$w^T x + b < 0 \quad \text{for} \quad y_i = -1$$

Hence $g(x) = w_0^T x + b_0$ gives the discriminant function which is the algebraic measure of distance of x from hyper-plane.

## C. Logistic Regression

Logistic Regression makes use of logistic function for classification of input datasets. Also called as sigmoid function, it is a S-shaped curve whose input can be any real number. It then maps the input values to a value between 0 and 1. It can be represented as follows:

$$\text{Sigmoid}(x) = \frac{1}{1+e^{-x}}$$



Logistic Regression can be considered as probabilistic linear classifier. It basically predicts the value of Y (dependent) variable for each of the X(independent) variable. The basic difference between linear regression and logistic regression is that instead of measure the Y variable directly, logistic regression measures the probability of obtaining a particular value of a variable.

## V. CONCLUSION

In this paper, we have studied about various machine learning algorithms. These algorithms can then further be used in the field of healthcare for early prediction of multiple diseases and epidemic. Using these algorithms multiple self learning systems can be developed which can aid doctors for better and easier diagnosis of multiple diseases. Decisions related to patient healthcare can be made for the scope of efficient outcomes.

## VI. REFERENCES

[1] Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, Lei Hua, "Data Mining in Healthcare and Biomedicine: A Survey of the Literature," Springer Science+Business Media, LLC 2011.

[2] P. K. Anooj, “Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules,” J. of King Saud Uni. Comput.and Inform. Sci., ELSEVIER, Vol. 24, pp. 27-40, 2012.

[3] Muhammad Zubair, Kibong Song, Changwoo Yoon,”Human Activity Recognition Using Wearable Accelerometer Sensors” Information & Communication Network Department, Korea University of Science and Technology

[4] Vinitha Dominic, Deepa Gupta, Sangita Khare, “An Effective Performance Analysis of Machine Learning Techniques for CardioVascular Disease”, Applied Medical Informatics. March 2015, Vol 36, No 1, pp: 23-32.

[5] Daniel Vieira, X-akseli, Oy Espoo, “Resource Frequency Prediction in Healthcare: machine learning approach ,” Aalto University School of Science, Department of Information and Computer Science, Espoo, Finland

[6] Min Chen, Yixue Hao, Kai Hwang , “Disease Prediction by Machine Learning over Big Data from Healthcare Communities”

[7] S. Patel and H. Patel, “Survey of data mining techniques used in healthcare domain,” Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60,March 2016.

[8] Anitha Ilapakurti, Jaya Shankar Vuppalapati, Santosh Kedari, Sharat Kedari, Chitanshu Chauhan, “ iDispenser - Big Data Enabled Intelligent Dispenser, ”Hanumayamma Innovations and Technologies Inc., 628, Crescent Terrace, Fremont, CA 94536

[9] Y. Hayashi and S. Yukita, “Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset,” Informatics in Medicine Unlocked, ELSEVIER, Vol. 2, pp. 92-104, 2016.

[10] S. Sa’di, A. Maleki, R. Hashemi, Z. Panbechi and K. Chalabi, “Comparison of data mining algorithms in the diagnosis of type II diabetes,” Int. J. on Comput. Sci. & App., Vol. 5, pp.1-12, October 2015.

[11] G.Huang, K.Huang, T.Lee, J. Tzu-Ya and Weng, “An interpretable rule based diagnostic classification of diabetic nephropathy among type 2 diabetes patients,” Huang et al. BMC Bioinformatics, Vol. 16, pp.55-65, 2015.

[12] Han, J. and Kamber, M. (2006), “Data mining: Concepts anf Techniques, Second Edition, Morgan Kauffmann Publishers, San Francisco

[13] Shashikant Ghumbre, Chetan Patil, and Ashok Ghatol, “Heart Disease Diagnosis using Support Vector Machine”