International Journal of Recent Trends in Engineering & Research
ISSN (ONLINE) : 2455 – 1457
IMPACT FACTOR : 4.101

A REVIEW OF MACHINE LEARNING ALGORITHMS IN HEALTHCARE

Preetha S[1], Abhishek Manohar[2], Adithi Aithal H M[3], Namana Y Tarikere[4]

[1,2,3,4] Department Of ISE, B.M.S. College Of Engineering, VTU, Bengaluru, India

Abstract—Machine Learning deals with study of computer algorithms which automatically improve the accuracy of predictive models through experience. It is a sophisticated technological trend that has various applications in every field such as Finance, Medical, Business, and Manufacturing etc. The field of Machine Learning has its roots in algebra, statistics, Medical Science and in Security. Machine Learning analyses, understands and recognizes a pattern in the data. We review different Machine Learning Algorithms used for detecting various diseases in this paper. The various Machine Learning algorithms help to build decision support systems.

Keywords— Machine learning, Disease prediction, Healthcare, Alzheimer's disease, Lung Cancer, Diabetes.

I. INTRODUCTION

In today's world of business and society, Machine Learning (ML) and associated innovations present have proven to be of tremendous help in the healthcare industry. These innovations are cable of transforming many areas of patient healthcare and administrative processes within insurer, payer and pharmaceutical organizations. ML is used to evaluate the significance of clinical parameters and their combinations for prognosis, e.g. for forecasting disease development, extracting medical information for outcome testing, preparing and supporting treatment, and for overall patient management.ML is often used for data processing, such as identification of data regularities by handling imperfect data correctly, interpretation of continuous data used in the Intensive Care Unit, and intelligent alarming resulting in accurate and efficient monitoring. The data in the healthcare field is enormous and complex. Machine Learning Techniques have proven to be helpful in analyzing and processing data to obtain the desired result. Research studies suggest that Artificial Intelligence and Machine Learning have a vital role in healthcare. Machine learning is a technique that statistically fits a model to data. An algorithm learns through the training data and then is applied on the test data. In healthcare, predicting what treatment best suits a patient according to the various patient attributes is one of the most common applications of Machine Learning.

Machine learning techniques have been applied in various tasks such as body scan, image recognition, detection of diseases, patient monitoring, enabling remote healthcare services etc. More precise analytics are obtained due to the automation of processes. With the help of evolving technologies such as cloud/edge computing, mobile computing and big data technologies, the application of Machine Learning methods in healthcare has been more effective and practical. A study on the applications of Machine Learning in healthcare thus proves to be of high importance. Three major diseases that are commonly seen in the society that can be detected, tested and analyzed by using various machine learning techniques have been chosen. The diseases chosen are Alzheimer's', Lung cancer and Diabetes.

II. LITERATURE SURVEY

Wang et al.[1] describes the transformation of cells from healthy to cancerous lung nodules by the method of wavelet transformation based on subtraction and decomposition algorithm. They applied image processing technique to recognize lung tissue information. The auto-detection of the tiny nodules which represent the information regarding early lung cancer is a major aspect. The newly developed ridge detection algorithm is used to identify and diagnose intermediate nodules accurately and allows detection of early-stage malignant lung nodules along with cure and prevents the death rate involved in surgery for benign cancer cells. The approach used is edge detection algorithm which is adaptive to perform in different environments and detect all the subsequent stages of processing in detecting cancerous nodules. The computerized tomography (CT) scanned image of maximum extension in gradient reflects the characteristic of use of edge detection algorithm. Observations made from the research and analysis is that the wavelet transformation algorithm successfully detected pulmonary solitary lung nodules that are unusual findings on a lung CT scan. Benign tumours are nodules with smooth edges whereas malignant tumour cells are with ragged or rough edges. The size of tumours also maps with the possibility of presence of malignant tumours in individual pulmonary lung nodules. The probability of cancer in every pulmonary nodule which were not specifically malignant or benign cells after obtaining density, color and growth characteristics were classified to evaluate and analyze patients for radiographic waves, UV rays and microwaves specific risk factors.

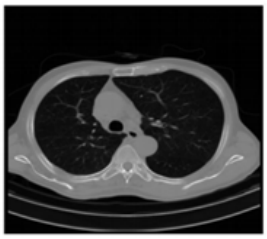
Amita Dessai et al. [2], proposed a lung cancer analysis and identification algorithm. The algorithm used morphological and mathematical operations for the purpose of separation and segmentation of the suspected region of affected lung tissue from which Haralick features are extracted. Haralick features are a set of features that are analyzed and derived from training huge data sets and are used for classification and filtering of cancerous cells by artificial neural networks algorithm. A huge collection of dataset is analyzed to extract most desirable features that are observed in patients suffering from lung cancer and the outcome of the analysis is represented by Haralick features. The amount and speed at which cancer spreads is based on the division of lung cancer into different stages. Various stages are: stage 1 - Cancer is identified and present in the lungs, stages 2 and 3 - Cancer is spread to the chest (with more invasive, larger and damaging tumors) and Stage 4 - Cancer has confined from lungs to other parts of the body.

The methodology used for the detection and classification are :

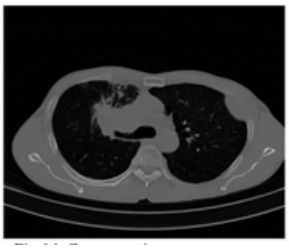
- Unprocessed CT Imaging - By data collections and analyzing cancerous from non-cancerous CT images
- Preprocessing - By reading and cropping lung CT image and obtaining application of median filter
- Segmentation - By filtering out lungs my max area and obtaining the density of the lungs mask by superimposing the images
- Feature extraction - By creating GLCM from images and calculating Haralick features from GLCM
- Classification - By using Feed forward Artificial neural networks

Techniques that are used in [2] to diagnose lung cancer are X-rays, CT scans, Magnetic Resonance Imaging, and Sputum Cytology. The analysis is focused on developing an automated technique to detect lung cancer. Median Filters are used to remove impulsive noise produced by the images successfully. Segmentation issues with partitioning of white matter in lungs were also considered for the analysis and prediction. Artificial neural networks made a perfect classifier with acceptable accuracy. The algorithm was successful in achieving an accuracy of 92% for the hospital database. Thus the algorithm aims at increasing the speed, effectiveness and accuracy of detecting

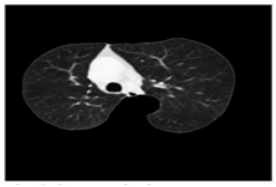
lung cancer in the system. It also effectively detects the early stage lung cancer. Figure 1 shows images of segmentation tissues, non-cancerous and cancerous images.



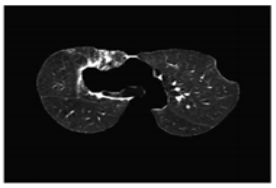
Non-cancerous image



Cancerous image



Initial segmentation tissue



Final segmentation tissue

Figure 1: Images of segmentation tissues, non-cancerous and cancerous images

Anuradha Thakare et al.[3] focused on design and development of an algorithm for diagnosis and identification of preliminary lung cancer from CT, PET images and X-Ray. Genetic Algorithm is used to analyze and optimize the results. The Genetic algorithm aimed to detect early stage cancerous nodules present in the lungs. The human interpretations and operations are very critical and time consuming. Hence to eliminate the inconvenience, Genetic Algorithm and Naïve Bayes Classifiers are applied to identify the various stages of cancer images accurately and swiftly. The proposed system resulted in 80 percentage accuracy in classification of the disease.

The genetic algorithm is through the exchange of genes between parents which happens in following steps:

- Randomly generating a population of M chromosomes
- Calculation of fitness value of each function in the chromosome population
- Repeating the procedure until large number of offsprings are produced
- Selecting a pair of chromosomes from current population using the fitness function value
- Producing offspring y, using crossover and mutation
- Replacing the current population with newly formed generation

Naive Bayes Algorithm classifies the cancerous from non-cancerous images and the results are optimized using Genetic algorithm. Input images are preprocessed Using Canny detection, the input images are preprocessed to detect strong edges and feature vectors are generated from the output obtained. Output provides accurate results and detects if the given image is cancerous or not.

In [4], computerized Tomography (CT) is 3D imaging modality which has been widely used for lung cancer screening and diagnostics. Most current machine learning based Computer Aided Diagnostic (CAD) research is focusing on NSCLC. These systems help to reduce the workload of radiologists significantly. Decrease in research on SCLC detection is because the image with SCLC looks very identical to non-cancerous cells. Wu, Qing, and Wenbing Zhao [4] proposed an artificial

neural-network based algorithm, which is referred to as entropy degradation method (EDM), to detect SCLC from CT images. The high resolution CT scans of training data and testing data were selected for analysis. Six of the data sets were from healthy lungs, and the remaining were the scans from patients with SCLC. When Five scans from each group were randomly selected to train the model, the algorithm achieved an accuracy of 77.8%. EDM is designed to transform the vectorized histogram of each training set into a score. Scores are transformed into probability through a logistic function and the difference between the label and transformation is calculated and fed back by a back-propagation stage called a score-probability policy. Results indicate to which group the testing data belongs.

Kancherla, et al.[5] proposes an early lung cancer detection methodology using nuclear segmentation features. The sputum samples from patients are identified with Tetrakis CarboxyPhenyl Porphine (TCPP) and fluorescent images of the samples are collected. TCPP is a porphyrin able to assist in finding cancer cells by increasing the count of low density lipoprotein coating present on the surface of cancer. Performance is measured by ML techniques for cancer identification and obtained an accuracy of 87% using 71 features related to shape, intensity and color by using the nucleus segmented features. By using the Seeded region growing segmentation method, nucleus segmentation is performed. After obtaining, preprocessing and segmentation of CT images, a group of pixels are clustered into larger regions. The steps involved are

- CT image enhancement
- Finding the seed point initially
- Find the threshold value for adding pixels to neighbouring cells
- Include neighboring pixels if they satisfy the criteria - seeded region growing method
- Repeat adding pixels and stop when neighboring pixels satisfy the criteria.

After performing the Seeded Region Growing Segmentation, the Results obtained are, If absolute difference between neighboring pixel and average of intensities in region is less than initial threshold and if absolute difference between neighboring pixel and intensity of current pixel is less than current threshold, lung cancer is predicted.

A basic framework for prior detection of Alzehimer's disease based on deep neural networks and clinical pertinent data was suggested by Ronghui et al.[6]. Functional connectivity of regions of the brain is calculated through the use of R-fMRI data. Functional magnetic resonance imaging reflects the spontaneous BO level variation when a person is not doing any particular task. It is an effective technique to find out how anatomically different and differentiated brain networks are interconnected. People can experience a medical stage between cognitive decline due to normal aging and cognitive decline due to dementia. This is known as Mild Cognitive Impairment. A targeted auto encoder network is built to differentiate between mild cognitive impairment and normal aging, the former being an early stage of Alzheimer's disease(AD). Medical history is also considered. Functional Magnetic resonance imaging scans as well as information about gender, age and genetics have been used for training the model and classifying the data. These brain networks have been constructed based on the association of R-FMRI signals. The brain networks have been constructed based on the correlation of R-fMRI signals, and then used as correlation coefficient data for training the deep neural network.

The method proposed in [7] unveils discriminative features of the brain network effectively and gives a reliable classifier for diagnosis of Alzheimer's disease. In comparison to conventional classifiers based on R-fMRI time series data, an improvement of about 31.21% in the accuracy of predictions is obtained by the deep learning technique. The standard deviation reduced by 51.23% in the best case indicates the model is more reliable and stable in comparison to the conventional methods. The method also implies that the amalgamation of deep learning and brain network is a

robust tool for detecting neurological diseases early. Based on this work, identical techniques can be used to detect other neurological illnesses. As an extended work authors suggested assessing this method on larger data sets and using it for detecting other neurological diseases.

The idea of unsupervised feature learning which uses Artificial Intelligence(AI) to discover attributes from data is used in [8]. A two stage method is proposed for smart diagnosis of Dementia. Features from the data are learnt through a two layer neural network in the first stage. Second stage involved using SoftMax regression to differentiate between the status of health based on learned features. Magnetic resonance imaging (MRI) Images were used to validate this method. The method attained decent accuracy and performed better than other traditional techniques for data sets of the brain image. It also significantly lessens the necessity for manual work and makes it easier to smartly identify for processing of big data, as the learning features are adaptive. A multi class and dual classification was conducted for Alzheimer's disease detection on Alzheimer's disease Neuroimaging Initiative data. The centre of interest is the hippocampal region of the brain. The dataset used is from OASIS. Various attributes of the hippocampus region of the brain are extracted for diagnosis of Dementia. Feature vector is generated using extracted features. Features such as age, gender, SES, MMSE are taken directly from the dataset. A gray level co-occurrence matrix is used to derive texture features, and other attributes such as area and shape are derived from seven moment invariants. The training set consists of 235 MRI scans, out of which 135 are CDR0 scans, 69 are CDR0.5, 29 are CDR1 and 2 are CDR2 scans. After training the network, it will be fit for classifying various stages of Dementia. A neural network consists of 3 layers. First one is the input layer, which is made up of 21 neurons in this case, as it is the size of the feature vector. Second one is the hidden layer, in this case, there are 4 hidden layers. The third and the last layer is the output layer. The output layer consists of four neurons, to categorize the subjects into one of the following categories - normal, mild, moderate and severe. The Error-back Propagation (EBP) in Artificial Neural Network (ANN) is used as the classifier for detecting various stages of AD. Average accuracy of 86.6% was obtained using this system.

In [9], prediction of a person's state in the coming times is done using physical, demographic and cognitive data collected from different times in the person's history. In time series analysis or prediction, weighted combinations of earlier values are used to forecast the next value. In general, time series models encode a mapping from an input space to the output, and time is not considered as one of the input dimensions. Random forest algorithm was used on this time series data. The time series data consisted of at least four points (of the patient's history), this number was determined while training the model. The average cross-validation accuracy obtained for this set was 82%. It was shown as to how machine learning methods can comprehend relationships among pairs of data points at various time periods for the detection of AD. The benefit of this method is that it can be effortlessly applied on data where data points are missing. It performs better than the SVM method. It exhibits the effectiveness of pairwise prediction methods and shows that it performs well in comparison with other approaches. Results obtained from this method also show that random forest is a successful option for these kind of classifications, as they take qualitative and quantitative inputs, they can be trained easily and their performance is good over a broad range of applications

R Rajeswara Rao et al.[10] contributes a comprehensive run-through of Prediction of AD by using various machine learning models. It also outlines the process of Brain image classification and gives an outline of the results obtained by other researchers for solving the same problem, i.e. Predicting Alzheimer's disease. Classification algorithms, Regression algorithms, Association algorithms and Clustering algorithms are analyzed in the paper. Support vector machines gave better results than other classifiers being analyzed. Further scope includes extracting a reasonable set of

features for detecting Dementia early and to reduce the number of in signification and repeating features.

Chen et al.[11], K-means clustering and decision tree algorithms are used to predict diabetes type 2. To implement the system, WEKA, software that works on java, is used and Pima Indian Diabetes Data (PIDD) set is used. The data is preprocessed in which the values that are missing are replaced by the mean. Data reduction involves removing the data that was incorrectly classified using the K-Means clustering algorithm using WEKA. Parameters such as Insulin, BP, BMI, Age and Diabetes History etc. are selected. Based on these features, classification into Diabetic and Non – Diabetic is done using the Decision Tree J-48 algorithm. The problem results in 2 classes, "Diabetes Positive" and "Diabetes Negative". The performance of the system is evaluated on the parameters Accuracy, sensitivity and Specificity. Based on Attribute information and performance evaluation, the patients suffering from Type 2 Diabetes and their symptoms are found out. The confusion matrix helps us understand the accuracy with which a classifier recognizes data belonging to different classes.

Development of a mobile health application (mHealth) to provide information on health related aspects through smartphones and other such devices was done in [12]. Through the application, the details of the individual such as body mass index (bmi), gender, age and inheritance are collected using survey questionnaires. To check if a person is prediabetic, diabetic or non-diabetic, the app uses the Naïve Bayes Classifier algorithm. This is done based on the features of the individual. The Naïve Bayes Classifier calculates the probability of a person belonging to a particular class. The algorithm is analyzed in two phases: Training and Testing. The implementation of the algorithm is done using JAVA using the IntelliJ IDE and an app is developed using Android Studio. This app implements the algorithm. Through this system, an individual can determine if they are diabetic or not. There is no need to visit a hospital. As this is a software tool, results are obtained immediately and thus the user is given enough time to control/prevent diabetes.

Decision Tree algorithm was used in [13]. The J48 decision tree algorithm chooses an attribute for splitting the data and examines the information gain which is normalized. The attribute with the maximum normalized information gain is used to make conclusions. The algorithm is then applied on subsets. Algorithm examines the real data collected from known hospitals and WEKA tool is used for implementation of algorithms. Data is collected and age is categorized into groups. An on-screen questionnaire is used to obtain data such as BP level, hunger and thirst frequency, food habits etc. from the user and machine learning algorithms are employed to learn the model. Data is pre-processed by replacing the missing values with the median. It is then partitioned into two sets for training and testing. Algorithm is applied on the data set and rules obtained are used to build the smartphone android application. Rules were obtained from decision trees. The additional IF THEN rules obtained from human heuristics are integrated to build the android application. The solution is provided through a mobile app, MobDBTest that predicts the probability of a person being diabetic and provides valuable information. Predictions are made using the learned model which is evaluated by constructing a confusion matrix.

In [14] Deep Learning and Machine Learning Techniques are used. The Support Vector Machines are considered to be an effective method for the classification of data. By finding the set of points lying on the edge of the class descriptors, the optimal separating hyperplane between classes is obtained. SVMs can handle data that are linearly separable and which are binary classified. A supervised algorithm based on classification of data is The Random Forest. There are a group of decision trees and every tree is characterized by an equal number of nodes and random variables. The results obtained from the trees are collected to obtain the end result representing the average

responses of all these trees. PIMA Indians diabetes dataset is used. The dataset consists of a number of instances and 8 features. The data set is so divided that 60% of the data is the training set and 40% is the testing set. 10% of the training data set is used for validation wherein the performance of the model is evaluated. Convolutional Neural Network is the algorithm used. This is an artificial neural network that makes a set of changes on the input to obtain the needed output. The input of the next block will be the previous layer's output. On applying the SVM and RF algorithms on the data, it was found that the RF algorithm is more efficient in the classification of data. Along with the prediction of diabetes, the study performs a differentiated analysis of machine learning and deep learning techniques.

A Method of analyzing data on diabetes by applying Machine Learning techniques in Hadoop Map Reduce Environment was proposed by Amani Yahyaoui et al.[15]. Various data mining techniques and machine learning algorithms are integrated using the predictive analysis model. In this model, the algorithms are implemented in a Hadoop-Map Reduce environment. A framework in java, Apache Hadoop, processes large amounts of data on a set of computers in a distributed manner. Hadoop provides distributed storage and processing of data. MapReduce framework works in two phases. In the Map phase, the conversion of input data into intermediate data occurs in the form of key value pairs. In Reduce phase, all the values associated with the key are integrated to obtain the final output. The data set used is Pima Indians Diabetes Data (PIDD). Missing values in the data set are replaced with the attribute mean using the MV1 classification clustering algorithm. The C4.5 decision tree algorithm is implemented to generate rules. Recognition of a pattern given a sequence of strings/values is termed as Pattern Discovery. Once the information gain ratio is measured, test characteristics at specific nodes of the tree can be selected. This type of measurement is known as attribute selection measure. Test feature for the present node is the attribute with highest information gain ratio. Different types of diabetes, complications associated and the suitable treatment for it is obtained. Based on this analysis, the system provides a solution for early diagnosis.

III. CONCLUSION

Machine learning (ML) techniques are important and play a key role in various business aspects. The Healthcare sector is expensive and has a lot of challenging issues. ML and its applications help in rectifying them. Through this paper, various ML techniques that can be used to predict diseases like Lung Cancer, Alzheimer's, and Diabetes are studied. Data processing leads to making important conclusions concerning the patient. Machine learning will be able to assist doctors and help them in being more efficient.

REFERENCES

I. Wang, Weixing, and Shuguang Wu. "A study on Lung cancer detection by Image processing." 2006 International Conference on Communications, Circuits and Systems. Vol. 1. IEEE, 2006.
II. Vas, Moffy, and Amita Dessai. "Lung cancer detection system using lung CT image processing." 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA). IEEE, 2017.
III. Kurkure, Manasee, and Anuradha Thakare. "Lung cancer detection using genetic approach." 2016 International Conference on Computing Communication Control and automation (ICCUBEA). IEEE, 2016.
IV. Wu, Qing, and Wenbing Zhao. "Small-cell lung cancer detection using a supervised machine learning algorithm." 2017 International Symposium on Computer Science and Intelligent Controls (ISCSIC). IEEE, 2017.
V. Kancherla, Kesav, and Srinivas Mukkamala. "Early lung cancer detection using nucleus segementation based features." 2013 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). IEEE, 2013.

VI. Ju, Ronghui, Chenhui Hu, and Quanzheng Li. "Early diagnosis of Alzheimer's disease based on resting-state brain networks and deep learning." IEEE/ACM transactions on computational biology and bioinformatics 16.1 (2017): 244-257.
VII. Razavi, Firouzeh, Mohammad Jafar Tarokh, and Mahmood Alborzi. "An intelligent Alzheimer's disease diagnosis method using unsupervised feature learning." Journal of Big Data 6.1 (2019): 32.
VIII. Raut, Arpita, and Vipul Dalal. "A machine learning based approach for detection of alzheimer's disease using analysis of hippocampus region from MRI scan." 2017 International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2017.
IX. Mishra, Siba Prasad, and Ananta Charan Ojha. "Evaluating the Impact of Naraj Barrage on Sedimentation of Chilika Lagoon using Random Forest."
X. G Stalin Babu , S N Tirumala Rao , R Rajeswara Rao, "Exploring of Classification Methods for Early Detection of Alzheimer's Disease",International Journal of Engineering and Advanced Technology (IJEAT)
XI. Chen, Wenqian, et al. "A hybrid prediction model for type 2 diabetes using K-means and decision tree." 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS). IEEE, 2017.
XII. Khan, Nabila Shahnaz, et al. "Diabetes predicting mhealth application using machine learning." 2017 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE). IEEE, 2017.
XIII. Sowjanya, K., Ayush Singhal, and Chaitali Choudhary. "MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices." 2015 IEEE International Advance Computing Conference (IACC). IEEE, 2015.
XIV. Amani Yahyaoui, Akhtar Jamil, Jawad Rasheed, Mirsat Yesiltepe,"A Decision Support System for Diabetes Prediction Using ML and DL Techniques"
XV. Amani Yahyaoui, Akhtar Jamil, Jawad Rasheed, Mirsat Yesiltepe, "Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop"

