1.

A comparative analysis of Nonlinear Machine Learning Algorithms is presented in this paper for Detecting Breast Cancer. [1]

**Motivations:** An abnormal multiplication of cells in the breast tissue causes breast cancer and it is the second leading cause of death among women worldwide. Early detection is the best way to increase the chance of treatment and reduce this cause of death. Machine Learning is an advanced computing technology which can significantly save time and reduce error. Machine Learning algorithms are applied to improve the analysis of medical data, early diagnosis and screening for breast cancer detection. Recently, multiple nonlinear machine learning algorithms are applied on medical datasets to perform predictive analysis about patients and their medical diagnosis like - using machine learning techniques to assess tumor behavior for breast cancer patients. This paper introduces a comparison between five nonlinear machine learning algorithms namely Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Gaussian Nave Bayes (NB) and Support Vector Machines (SVM).

**Objective:** The objective of this paper is to evaluate the performance in classifying data with respect to efficiency and effectiveness of each algorithm in terms of classification test accuracy, precision, and recall.

**Proposed Methodology:**

In this paper, the performance comparison between the five nonlinear machine learning algorithms is conducted on the Wisconsin Breast Cancer Diagnostic (WBCD) dataset for the classification. At first, the breast image database is loaded. Then the features are extracted and the classification model is trained and used for prediction of benign and malignant. Benign cases are considered noncancerous which is non-life threatening. Malignant cancer starts with abnormal cell growth and might rapidly spread nearby tissue which is life-threatening. The five different machine learning algorithms which are applied for the classification are briefly described below-

Multilayer Perceptron (MLP):

Multilayer Perceptron algorithm is a neural network which is connecting multiple layers in a directed graph. The general model of MLP with an input layer, hidden layer, and an output layer. Each node has a nonlinear activation function apart from the input nodes. The simplified sigmoid function, $\varphi(x) = 1 / (1 + e^{-x})$ is in hidden layer. And the sigmoid function $\varphi_0(x) = \varphi(x)$ is in output layer. This MLP model is implemented for the classification of breast cancer dataset . From the

input of 9 attributes of the database, the MLP has hidden layer with five neurons and the output layer generate two outputs for two class of cancer classification as benign and malignant.

K-nearest neighbors (KNN):

K-nearest neighbors (KNN) algorithm is used to feature similarity for predicting the values of new datapoints and the new data point is assigned a value. This value is assigned based on how closely it matches the data points in the training set. The distance between test data and each row of training data is calculated with the help of any of the method like Manhattan, Euclidean, Hamming distance. The most commonly used method is Euclidean. After calculating the distance, the value is sorted in ascending order. Then the top K rows from the sorted array are chosen. And then a class is assigned to the test point based on most frequent class of these rows.

Classification and Regression Trees (CART):

CART algorithm consists of three steps and the first step is to splitting the attributes. An overgrown tree is built which closely describes the training set. This tree is called the maximal tree. The tree is grown using a binary split-procedure. The second step is stopping rules for deciding when a branch is terminal and can be split no more. A series of smaller subtrees are derived from the maximal trees during this procedure. The overgrown tree shows that overfitting is being pruned. And finally, the tree with the optimal size is selected using a cross-validation procedure.

Gaussian Naive Bayes (NB):

This algorithm is used when the features have continuous values or all the features are following a Gaussian distribution. Naive Bayes classifier calculates the probability of an event by following some steps -The first step is to calculate the prior probability for given class labels. The second step is to find likelihood probability for each class with each attribute. The third step is to assign all these value in bayes formula and calculate posterior probability. And the last step is to see which class has a higher probability and give the input belongs to the higher probability class.

Support Vector Machine (SVM):

SVM algorithm is used to find a hyperplane in an N-dimensional space where N is the number of features that distinctly classifies the data points. The main focus of this algorithm is to find a plane that has the maximum margin (the maximum distance between data points of classes). In SVM, the output of a linear function is taken. The linear hyperplane is defined as ax1+bx2. So,the goal is to find a, b and c such that $ax1+bx2 \leq c$ for class 1 and that $ax1+bx2 > c$ for class 2.

**Data collection:** The data used in this work is the Wisconsin Breast Cancer Diagnostic dataset which is available at the UCI Machine Learning Repository .

**Dataset Description :** The dataset has 569 instances and 32 attributes. The dataset has a number of diagnosis of lumps and masses that were found in the patients. The tumor or lump is either classified as malignant (M) or benign (B) based on the diagnosis .The class distribution of the samples is such that 357 are benign and 212 are malignant.

**Dataset Preprocessing :** For this work, the data is standardized and the standardized value of a feature is called a Z score. Z score is calculated using a formula.

The contribution of this paper is to presenting machine learning algorithms with accuracies, precisions and recalls for the dataset to find out algorithm's behavior. A heat map plot of the correlations between the features has been shown which indicates how related the changes are between the two features in the dataset. If two features change in the same direction, then they are positively correlated. If two features change in opposite directions, then they are negatively correlated. It shows which features have a high correlation with each other which is important to know about some machine learning algorithms that can have poor performance if there are highly correlated input features in the data.

**Results:**

The dataset is splitted into two - 80% for training and 20% for testing. The accuracy, precision and recall is calculated to evaluate the performance of the five classification algorithms. Accuracy is the ratio of the number of instances cases correctly classified divided by the total number of instances cases. It has been shown that MLP has the highest prediction accuracy of 99.12%. Precision is the ratio of correct positive results divided by the total number of all predicted positive observations. It has been shown that MLP has 99% precision which is the highest value among all classifiers. Recall is the ratio of correct positive results to all observations in actual class. It has been shown that MLP has 99% recall which is the highest value among all classifiers. So, the results shows that MLP classifier has the best performance in terms of accuracy, precision, and recall. So, MLP model is the best classifier among the five proposed classifiers for classifying a tumor as benign or malignant.

This paper will help our project to analyze and compare classification accuracy, precision and recall of different machine learning algorithms on our dataset.

[1] A. Al Bataineh, "A comparative analysis of nonlinear machine learning algorithms for breast cancer detection," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 3, pp. 248–254, 2019, doi: 10.18178/ijmlc.2019.9.3.794.

2.

A comparative analysis of various machine learning algorithms is presented in this paper for detecting Dementia. [2]

**Motivations:** Dementia is a collective term used to describe various symptoms which can occur when few groups of cells of brain stop working in an appropriate manner. The clinical diagnosis of dementia is based on neurological examination and neuropsychological tests. The manual diagnosis which is time intensive and slow process that requires information like - neuropsychological test score,knowledgeable informant reports and so on. The accuracy and efficiency of the diagnosis are found out by the professional level of the practitioner. It will be a much more difficult task for classification and detection of dementia in several remote areas for lacking professional personnel. Machine Learning is an advanced technology which can analysis medical data and automatically make the diagnostic decision. In order to diagnose the dementia with expert systems, identifying using machine learning algorithms such as Multilayer Perceptron, J48, Naïve Bayes,Random Forest. To reduce the number of neuropsychological tests, various machine learning algorithms are used to classify dementia patients. In this paper, the following machine learning classification algorithms are applied for detecting dementia-J48, Naïve Bayes,Random Forest and Multilayer Perceptron.

**Objective:** The objective of this paper is to analyze the algorithms in terms of classification test accuracy to find out which algorithm perform best for the detection of Dementia.

**Proposed Methodology:** The dataset is collected from the OASIS-Brains.org. The OASIS (Open Access Series of Imaging Studies) dataset contains two types of data - Cross sectional MRI data and Longitudinal MRI data of non-demented and demented older adults. The algorithms are applied to both the datasets using WEKA tool. The attributes that included in the OASIS dataset are age, sex, education,socioeconomic status , mini-mental state examination , clinical dementia rating, atlas scaling factor, estimated total intracranial volume, and normalized whole-brain volume. The steps are briefly described below involved in this work-

**Data Collection:**

The dataset is collected from the oasis-brains.org. The dataset includes 373 records in longitudinal data and 416 subjects in cross-sectional data.

**Data Pre-processing:**

The missing values are filled up by using the average values.

**Feature Selection:**

For eliminating the redundant attributes ,CFSSubsetEval is used.

**Classifiers:**

J48:

C4.5 (J48) is used to generate a decision tree. The classification process is formed by using the binary tree and applied to all the tuples in the database.J48 is an extension of ID3 algorithm.

Naïve Bayes:

Naive Bayes is a classification technique based on Bayes Theorem which is used to calculate the set of probabilities by counting the value and frequency of values in a given set of data.

Random Forest

The tree predictors are combined in this algorithm. Each tree depend on the values of random vector sampled independently and have the same distribution for all of the tree.

Multi-layer Perceptron:

Multi-layer Perceptron is a supervised learning algorithm that includes three layers of nodes using nonlinear activation function. This algorithm consists of interconnected neurons transferring information to each other like the human brain. Each neuron is assigned a value and the network can be divided into three main layers.

The contribution of this paper is to presenting machine learning algorithms with accuracies to find out algorithm's behavior.

**Results**

The classification accuracy is obtained and analyzed. To find out the classification accuracy, the percentage of number of correctly classified samples divided by the total number of samples obtained is calculated. J48 algorithm is performing best among all the algorithms for the detection of Dementia. J48 algorithm is performing 99.52% classification accuracy and after attribute selection accuracy for Oasis Cross Sectional Data. Random Forest algorithm has 75.96% accuracy which is the least of all algorithm in case of oasis cross sectional data. J48 algorithm is performing 99.20% classification accuracy and 98.66% after attribute selection accuracy for oasis longitudinal data. Multilayer Perceptron algorithm is performing 74.53% which is the least of all in the case of Oasis Longitudinal Data.

This paper will help our project to analyze algorithm's behavior and compare classification accuracy of different machine learning algorithms on our dataset.

[2] D. Bansal, R. Chhikara, K. Khanna, and P. Gupta, “Comparative Analysis of Various Machine Learning Algorithms for Detecting Dementia,” *Procedia Comput. Sci.*, vol. 132, pp. 1497–1502, 2018, doi: 10.1016/j.procs.2018.05.102.