DETECTION OF ALZHEIMER'S DISEASE USING GRADIENT BOOSTING ALGORITHM

Abarna S[1], Pranesha R A[2] , Hafsana Fathima A R[3] , Priyadharshini K[4], Karthikeyan T[5]

[1-4](UG Student, Department of CSE, Knowledge Institute of Technology, Salem)
[5](Assistant Professor, Department of CSE, Knowledge Institute of Technology, Salem)

ABSTRACT: Alzheimer disease is an irreversible brain disorder which results in an impairment in the ability to perform daily activities. Memory loss is one of the primary factor of Mild Cognitive Impairment(MCI). The phase of MCI wind-up with a marked decline in cognitive function. The inability to find other "hidden" indicators of Alzheimer's factor that could be revealed by the routine blood tests, Advanced machine learning method called Correlation Explanation (CorEx) could be implemented in order to find something when you don't know of what you are looking for? Machine learning is a scientific study of algorithms and statistical models that the computer could perform the specific function without using any explicit guidelines. Evaluation of threat and early diagnosis of AD is the problem identified thus in our proposed system helps to predict the alzheimer's disease earlier with the help of algorithms in supervised learning and semi-supervised learning. It also generates a report which contains accuracies of algorithm we have used. Evaluation of threat and early diagnosis of AD is a key to prevention or to decrement the progression of the disease, whereas in the preceding research where based on risk factors for AD generally utilizes statistical comparison tests or incremental selection with regression model. The accuracy, sensitivity, and specificity achieved using the proposed method are superior to those obtained by various conventional AD prediction method.

Keywords: Alzheimer's disease, Machine learning, Gradient boosting algorithm and other supervised machine learning algorithms.

I. INTRODUCTION

Health care is an organized provision of medical care to individuals or community. Health care is performed by health professionals in allied health fields. Doctors and physician associates are a part of these health professionals. It includes work done in providing primary care, secondary care, and tertiary care, as well as in public health. All the medicines and treatment would be carried by the Healthcare, For example the disease which has got a rapid awareness and threat among the people is COVID-19. All the preventive measures, treatments, allocation of doctors and hospitals are done in the responsibility of the health care. In this we concentrate on the brain disorder which is known as Alzheimer's disease is a irreversible brain disorder, through supervised learning we are going to produce the graph which conveys about the accuracy by implementing the classification algorithms with the help of confusion matrix, by which we can comparatively decide which algorithm would yield higher accuracy and specificity.

II. RELATED WORKS

AlexandreSavio et.al had studied defective diagnosis can only be made after a post-mortem study of the brain tissue. The Support Vector Machine (SVM) either with linear or non-linear kernels are the state of the art tobuild up classification and regression systems. This will give results on the global feature vectors, the simple voting of independent classifiers based on statistical significance of VBM, the weighted combination of individual cluster SVM based on training errors, and an adaptive boosting strategy for combining classifier. GLM design without covariates can detect subtle changes between AD patients and controls that lead to the construction of SVM classifiers with a discriminative accuracy of 86% in the best case

To overcome the optimization problems the twin based SVM algorithm are demonstrated by the SaruarAlam et.al .A key advantage of this technique is its non invasiveness[3]. A primary focus of these studies was the large dimensionality of extracted features and the identification of disease signatures among them where the most discriminative information of the said diseases exists. Tomar and Agarwal reviewed several types of twin SVM algorithms, their optimization problems, and their applications. Our proposed detection method for the ADNI dataset yielded an accuracy of 92.65±1.18% with high sensitivity and specificity.

Linda Mary et.al suggested the Naïve Bayes and decision tree for early diagnosis. Alzheimer's disease is one of the types of the dementia which contribute to 60-80% of mental disorders, Early diagnosis of AD is important for the progress of more powerful treatments.Decision trees are commonly used in operations research, especially in decision analysis, to help identify a strategy which most likely reaches a goal, but they are also a popular tool in machine learning[5].Decision Tree algorithm on the resultant dataset that helps us to classify whether the patient is suffering from Alzheimer's Disease or not.Naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable[4].NaiveBayes algorithm is used on the resultant dataset to detect the stage of the disease the person is currently in.

To predict conversions between clinical categories, with a cross validation, the author Piers Johnson et.al identified the Genetic Algorithm[6]. Outcomes of these methods tend to emphasize single risk factors rather than a combination of risk factors. Genetic Algorithm(GA) can be useful and efficient for searching a combination of variables for the best achievement (Eg.Accuracy of diagnosis).In this study, a GA was used to select one or more sets of neuropsychological tests (features) which can predict AD progression with high accuracy and a logistic regression (LR) algorithm was used to build prediction models. In GA, the potential solutions compete and mate with each other to produce increasingly filter individuals over multiple generations. Each individual in the population (called genome or chromosome) represents candidate solution to the problem.

In this research paper the Jordan hardoop performed the classification of disease with and without imaginary with gradient boosting. Early detection of AD enables family planning and may reduce costs by delaying long-term care. Accurate, non-imagery methods also reduce patient costs. The Open Access Series of Imaging Studies (OASIS-1) cross-sectional MRI data were analysed. A gradient boosted machine (GBM) predicted the presence of AD as a function of gender, age, education, socioeconomic status (SES), and a mini-mental state exam (MMSE).The GBM achieved a mean 91.3% prediction accuracy (10-fold stratified cross validation) for dichotomous CDR using socio-demographic and MMSE variables. MMSE was the most important feature. ResNet-50 using image generation techniques based on an 80% training set resulted in 98.99% three class prediction accuracy on 4139 images (20% validation set) at Epoch 133 and nearly perfect multi-class predication accuracy on the training set (99.34%). Machine learning methods classify AD with high accuracy. GBM models may help provide initial detection based on non-imagery analysis, while ResNet-50 network models might help identify AD patients automatically prior to provider review.

III. PROBLEM IDENTIFICATION

Alzheimer's disease is a kind of dementia that would cause a severe memory loss and also causes brain cells death. This disease could be affected to every age group individuals, the regular monitoring and track of the patients are more important. Thus the problem identified was that to provide an earlier diagnosis as in the worst case the patient may also die due to delay in detection of Alzheimer's disease. Machine learning techniques could be implied for the earlier diagnosis of the disease as this concentrate more on providing the accuracy and efficiency of what solution we are seeking for. So supervised learning is used for the detection of this Alzheimer's disease as it uses a labelled data analyse , all these data could be classified with the help of N number of algorithms. The accuracy has been measured under the algorithm such as Support vector machine (79.56%), Logistic Regression (85.01%), Decision Tree (80.15) and Gaussian Naive Bayes (76.28).These where the Algorithms which was implemented so far in all the research papers but while comparing to all these algorithm, Boosting algorithms could be introduced. Boosting is an ensemble meta-algorithm in machine learning for the purpose of reducing bias, variance in supervised learning. This algorithm is one of the machine learning algorithms which is used for converting weak learners to strong ones. If an algorithm achieves hypothesis boosting rapidly, it is simply known as "boosting". This algorithm is 81.75% accurate. While implementing with Adaboost algorithm also the accuracy was 88.78%. Even though the accuracy level obtained in boosting algorithm is high when compared to other algorithms but we expect to obtain more accuracy so that the early diagnosis would be more easier.

IV. PROPOSED MODEL

In our proposed system it helps to predict the alzheimer's disease earlier with the help of algorithms in supervised learning and semi-supervised learning, thus the earlier prediction may be helpful to the doctors and also patients to get cured and medicated. It also generates a report which contains accuracies of algorithm we have used. The processing of the data is carried out by Four modules namely,

- Data Cleaning
- Covariance Matrix
- Evaluating Models
- Predicting with Gradient Boosting

The detailed architecture Fig.1 described the processing of the data which is fetched from the patients report.

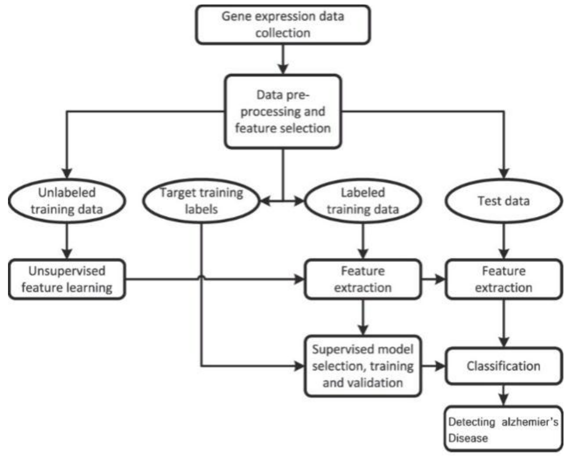

Gene expression data collection
Data pre-processing and feature selection
Unlabeled training data
Target training labels
Labeled training data
Test data
Unsupervised feature learning
Feature extraction
Feature extraction
Supervised model selection, training and validation
Classification
Detecting alzhemier's Disease

Fig. 1 Proposed Architecture of detecting alzheimer's disease

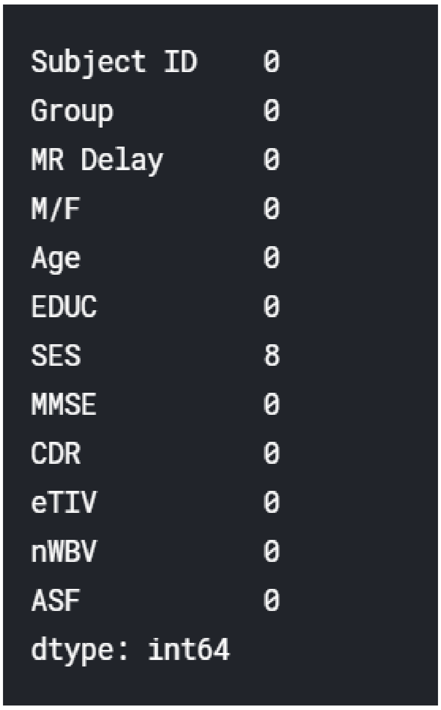
A.DATA CLEANING

Data cleaning plays a significant role while processing a large number of datasets from the mental health survey. When the data is used with the invalid or null data the generating of the final results becomes crucial thus all the irrelevant, inaccurate data is removed. Data cleaning may be performed interactively with data wrangling tools, and as batch processing through scripting. The data sets are cleansed to get high quality of data from the available data sets. The pseudo code of the missing data is given as below,

```
df_dropna = df.dropna(axis=0, how='any')
pd.isnull(df_dropna).sum()
df_dropna['Group'].value_counts()
df.isna().sum()
```

Pseudo code for noise removal

Data Encoding-Encoding Data also plays a vital role i.e, the Fig 2; with the help of this we can achieve data into an equivalent numerical format. **Label Encoder** is used to perform the encoding task, the label encoder class from the sklearn library will be helpful to fit and transform the data into a new encoded data.



```
Subject ID    0
Group         0
MR Delay      0
M/F           0
Age           0
EDUC          0
SES           8
MMSE          0
CDR           0
eTIV          0
nWBV          0
ASF           0
dtype: int64
```

Fig. 2 Data Encoding

B.COVARIANCE MATRIX

The data which is fetched from the Data cleaning is made as an matrix with i and j elements where the i and j position describes the covariance between the i-thand j-th elements of a random variables with this we can generate the entries of the covariance matrix is in the form of square matrix mentioned in Fig 3 which is denoted by,

- Ci,j=σ(xi,xj)Ci,j=σ(xi,xj) where C∈Rd×dC∈Rd×d.
- d*d describes the dimension or number of random variables of the data.
- Each element of this vector is called as scalar random variable these elements has either a finite number of observed empirical values or a finite or infinite number of potential values. By using the correlation between the vast set of data and variables is defined.

Seaborn-The Seaborn is Python data matplotlib based data visualization. . It is data set-oriented plotting function which operates on data frames and arrays

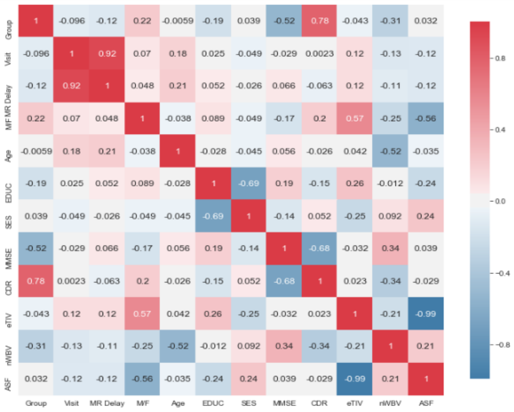


Fig. 3 Covariance matrix for variability comparison

Data relationship chart-The data relationship chart is generated with the help of the age of the patients i.e, Fig 4,5,6 who comes for the treatment, the chart is constructed with the two parameters

- **Treatment 0-** People who haven't taken the treatment.
- **Treatment 1-**People who undergoes treatment.

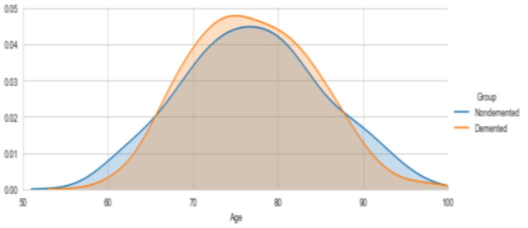


Fig. 4 Probability of Dementia

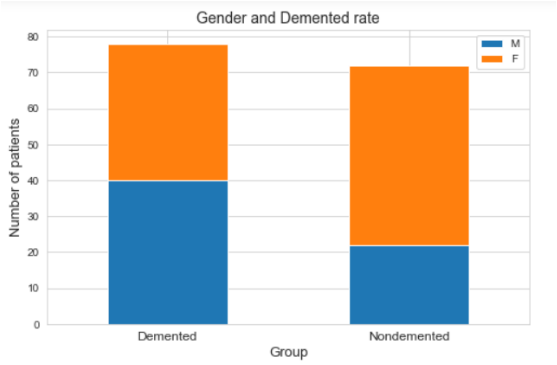


Fig.5Separated by Demented and Non demented

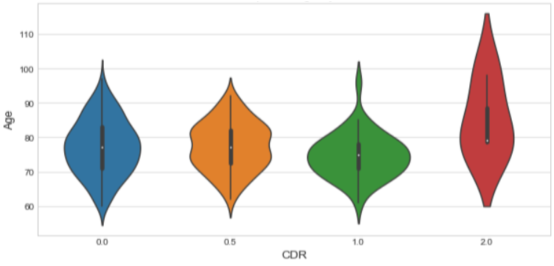


Fig. 6 Demented rating of AD patients

C. EVALUATING MODELS

The classification model will evaluate the following steps:

1. **Classification accuracy**: percentage of correct predictions

2. **Null accuracy**: accuracy that could be achieved by always predicting the most frequent class

3. **Percentage of ones**

4. **Percentage of zeros**

5 **.Confusion matrix:** Table that describes the performance of a classification model

- False Positive Rate
- Precision of Positive value
- **AUC:** is the percentage of the ROC plot that is underneath the curve

90-1 = excellent (A) 80-90 = good (B) 70-80 = fair (C)

60-70 = poor (D) 50-60 = fail (F)

The Fig 7 defines the evaluation models and parameters that is required for algorithms to define the confusion matrix.

Classification report

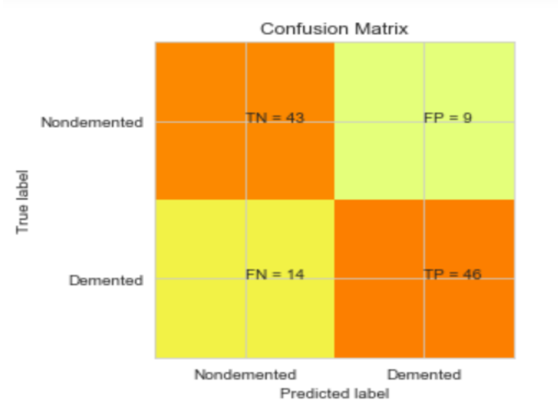
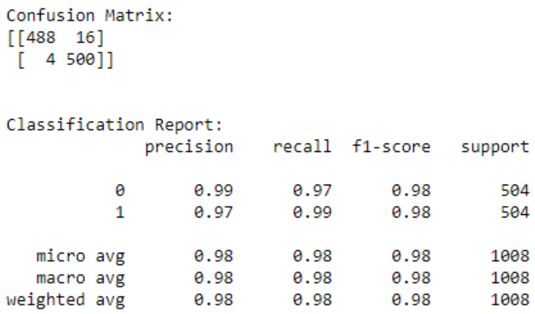


Fig. 7 Confusion Matrix

C.1 VOTING CLASSIFIER

Voting is one of the simplest way of combining the predictions from multiple machine learning algorithms Hard voting is the simplest case of majority voting. In this case, the class that received the highest number of votes Nc(yt) will be chosen. Here we predict the class label y via majority voting of each classifier. In soft voting, the probability vector for each predicted class (for all classifiers) are summed up and averaged. The winning class is the one corresponding to the highest value (only recommended if the classifiers are well calibrated).The accuracy of voting classifier is 97.33% is shown in Fig 8.



```
Confusion Matrix:
[[488  16]
 [  4 500]]


Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.97      0.98       504
           1       0.97      0.99      0.98       504

   micro avg       0.98      0.98      0.98      1008
   macro avg       0.98      0.98      0.98      1008
weighted avg       0.98      0.98      0.98      1008
```

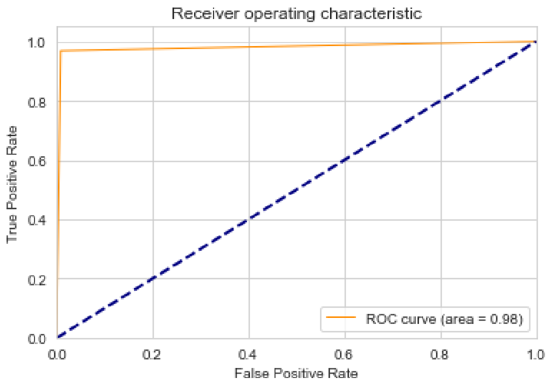
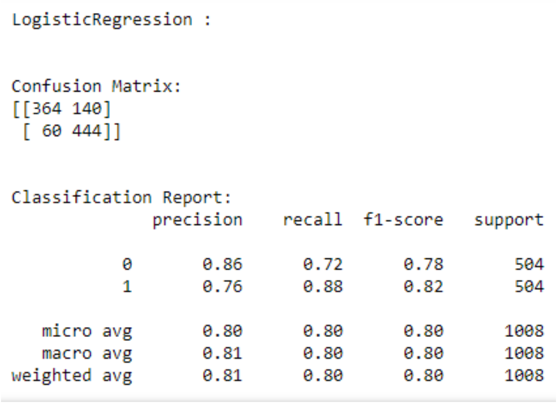


Fig 8 Voting Classifier

C.2 LOGISITC REGRESSION

Logistic regression is a technique used in machine learning from the field of statistics[6]. This method is used for binary classification problems. This algorithm is based on predictive analysis which is used to describe data and also it explains the relationship between one dependent binary variable and one or more nominal or ratio-level independent variables. This algorithm gives efficiency of 80.15% is shown in Fig 9.



```
LogisticRegression :


Confusion Matrix:
[[364 140]
 [ 60 444]]


Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.72      0.78       504
           1       0.76      0.88      0.82       504

   micro avg       0.80      0.80      0.80      1008
   macro avg       0.81      0.80      0.80      1008
weighted avg       0.81      0.80      0.80      1008
```

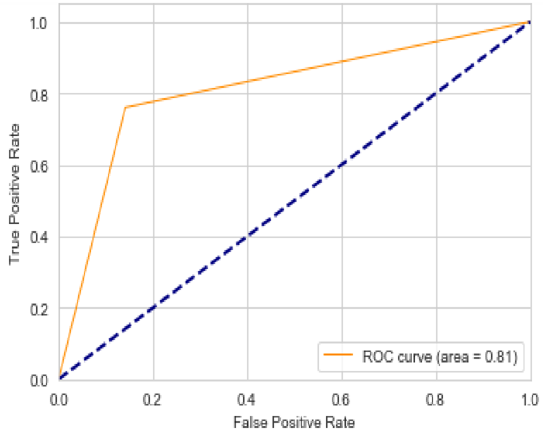
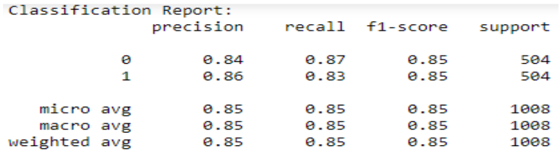


Fig 9 Logistic Regression

C.3 DECISION TREE

Decision tree is a type of supervised machine learning where the data is continuously split according to a certain parameter[9]. The tree can be explained by two entities, namely decision nodes and leaves[3]. The leaves are the decisions or the final outcomes. Decision tree classifier given in Fig 10 which gives an accuracy of 85.01%.



```
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.87      0.85       504
           1       0.86      0.83      0.85       504

   micro avg       0.85      0.85      0.85      1008
   macro avg       0.85      0.85      0.85      1008
weighted avg       0.85      0.85      0.85      1008
```

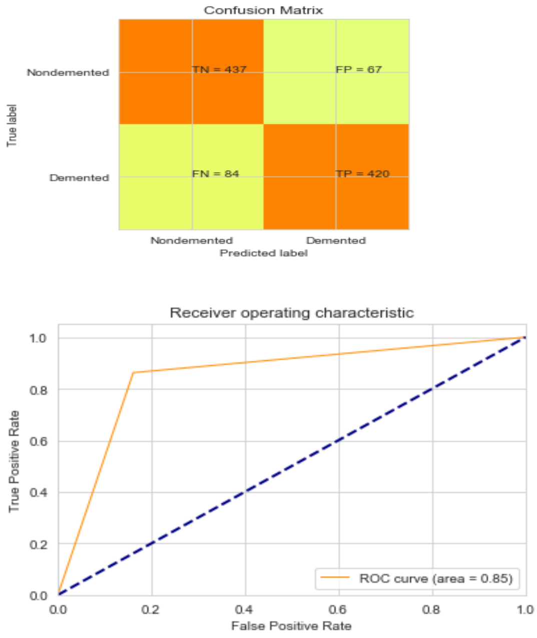
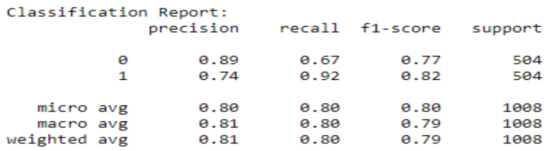


Fig 10 Decision tree

C.4 SUPPORT VECTOR MACHINE

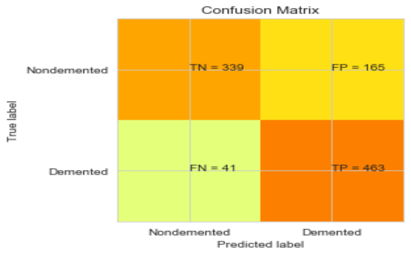
Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression[7].But generally, they are used in classification problems. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH)[2]. To conduct the first support vector machine (SVM)-based study comparing the diagnostic accuracy of T1-weighted magnetic resonance imaging (T1-MRI).Support vector classifier is shown in Fig 11, gives an accuracy of 79.56%.



```
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.67      0.77       504
           1       0.74      0.92      0.82       504

   micro avg       0.80      0.80      0.80      1008
   macro avg       0.81      0.80      0.79      1008
weighted avg       0.81      0.80      0.79      1008
```



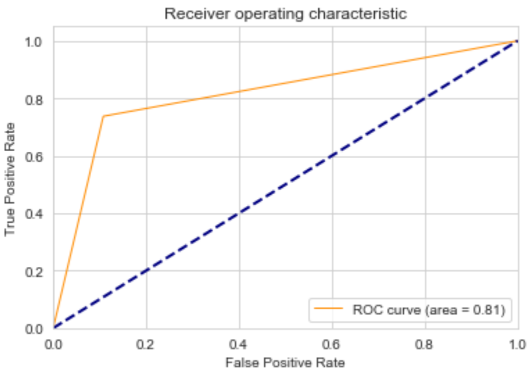
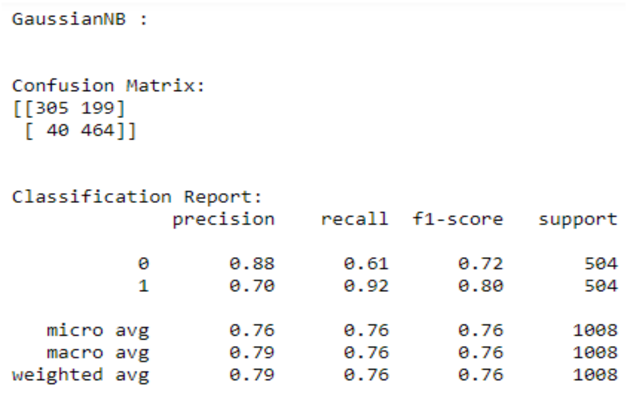


Fig 11 Support Vector Machine

C.5 GUASSIAN NAIVE BAYES

Naive Bayes is a kind of classifier which uses the Bayes Theorem. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class[8]. A frequency table for eachattribute iscreated and the likelihood of each feature is calculated is described in the Fig 12.Based on the likelihood, the conditional probabilities for each classes is determined, and the class with the maximum conditional probability is considered. The guassian naive bayes gives an accuracy of 76.28%.



```
GaussianNB :


Confusion Matrix:
[[305 199]
 [ 40 464]]


Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.61      0.72       504
           1       0.70      0.92      0.80       504

   micro avg       0.76      0.76      0.76      1008
   macro avg       0.79      0.76      0.76      1008
weighted avg       0.79      0.76      0.76      1008
```

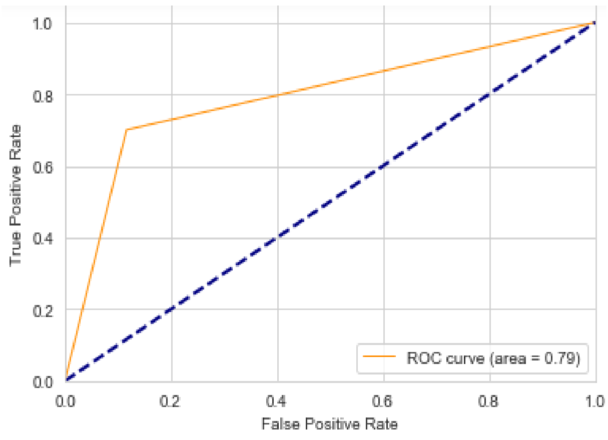
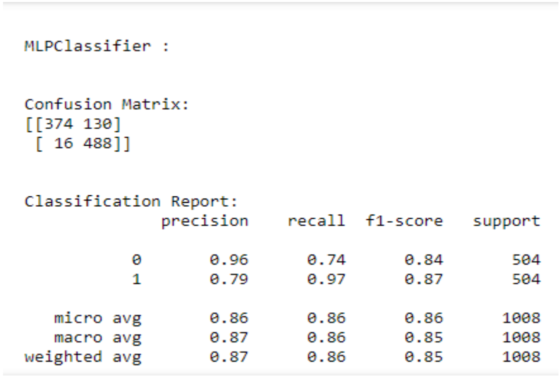


Fig 12Guassian Naïve Baise

C.6 MULTILAYER PERCEPTION

A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN).MLP utilizes a supervised learning technique called back propagation for training. Its multiple layer and non-linear activation distinguish MLP from a linear perceptron[10]. Supervised learning technique called backpropagation for training. It can distinguish data that is not linearly separable. The MLP and ANN are used for classification of Alzheimer's Disease and Parkinson's disease (PD) subjects. The MLP classifier in Fig 13gives an accuracy of 91.07%.



```
MLPClassifier :


Confusion Matrix:
[[374 130]
 [ 16 488]]


Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.74      0.84       504
           1       0.79      0.97      0.87       504

   micro avg       0.86      0.86      0.86      1008
   macro avg       0.87      0.86      0.85      1008
weighted avg       0.87      0.86      0.85      1008
```

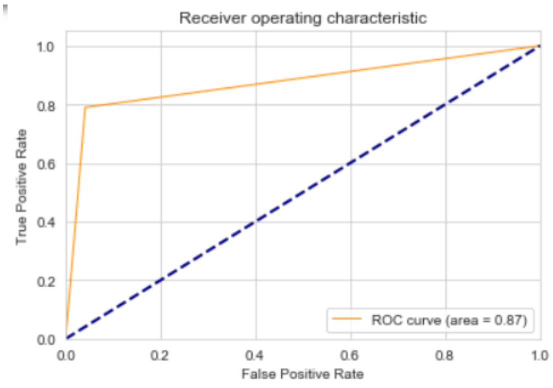
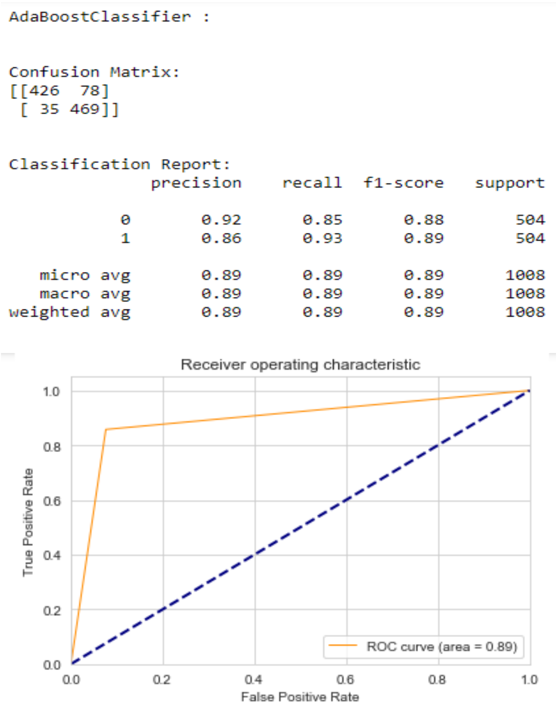


Fig 13 MLP

C.7 BAGGING

Bagging which is also known as Bootstrap aggregating, is an ensemble meta-algorithm in machine learning. This algorithm provides the stability and high efficiency in accuracy of machine learning algorithms which has been used in statistical classification and regression. This leads in decreases variance and also helps in avoiding the over fitting. AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners. These are models that achieve accuracy just above random chance on a classification problem[1].Adaboost strategy applied to the SVM built on the feature vectors. The selection are performed from the Open Access Series of Imaging Studies (OASIS) database, which is a large number of subjects compared to current reported studies. Results aremoderately encouraging from the Fig 14, as we can obtain up to 88.78% accuracy with the Adaboost strategy in a 10-fold cross-validation.



```
AdaBoostClassifier :


Confusion Matrix:
[[426  78]
 [ 35 469]]


Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.85      0.88       504
           1       0.86      0.93      0.89       504

   micro avg       0.89      0.89      0.89      1008
   macro avg       0.89      0.89      0.89      1008
weighted avg       0.89      0.89      0.89      1008
```

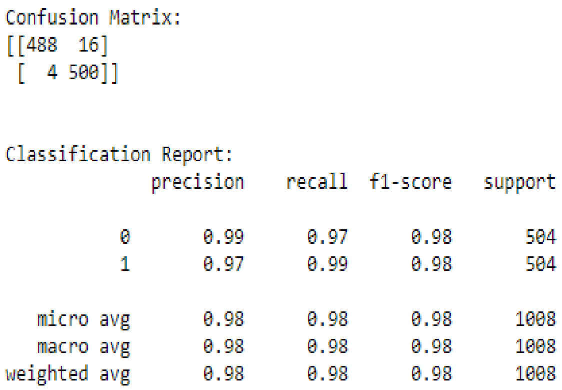
Fig 14Adaboost

D.PREDICTING WITH GRADIENT BOOSTING

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of ensemble of weak prediction models, typically decision tree. Specifically, our method first adopts Gradient Boosting Decision Tree (GBDT) to learn the 1H-MRS biomarkers of EOAD patients, which are then used to construct the final classifier for Alzheimer diagnosis. To validate our proposal, we have conducted comprehensive experiments for evaluation and the experimental results clearly demonstrate the effectiveness of our method. Hence, the gradient boosting is the best algorithm to predict the early alzheimer'sdisease, ig 15 provide an highest accuracy of 97.22% from the above algorithms.

	SVM		Decision Tree		Adaboost		MLP		Guassian NB		Logistic Regression		Gradient Boosting	
	0	1	0	1	0	1	0	1	0	1	0	1	0	1
Precision	89%	74%	84%	86%	92%	86%	96%	79%	88%	70%	86%	76%	**99%**	**97%**
Recall	67%	92%	87%	83%	85%	93%	74%	97%	61%	92%	72%	88%	**97%**	**99%**
F1-Score	77%	82%	85%	85%	88%	89%	84%	87%	72%	80%	78%	82%	**98%**	**98%**

Fig 15 Precision, Recall, F1-Score of Demented represented as 1 and Non-demented represented as 0, Gradient Boosting gives more accuracy while comparing with other classification report



```
Confusion Matrix:
[[488  16]
 [  4 500]]


Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.97      0.98       504
           1       0.97      0.99      0.98       504

   micro avg       0.98      0.98      0.98      1008
   macro avg       0.98      0.98      0.98      1008
weighted avg       0.98      0.98      0.98      1008
```

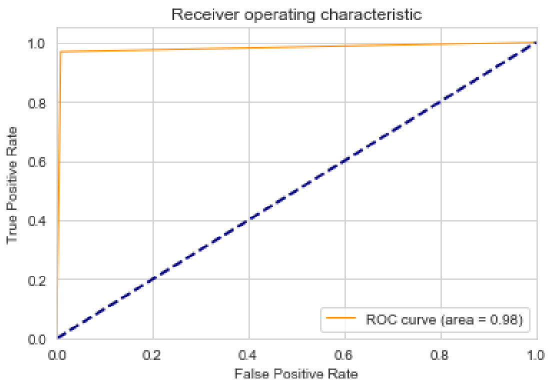


Fig 16 Gradient Boosting

V. RESULT AND DISCUSSION

The implementation of the proposed solution begins with installation of anaconda software. This process is followed by launching Jupyter notebook which helps to import the certain necessary packages i.e, pandas, numpy, sklearn etc. After importing all the packages, various machine learning are implemented for identifying an algorithm with high accuracy. The algorithm which is found to be more accurate is embedded with GUI (Graphical User Interface) backend for database connectivity.

In our proposed system we have obtained a better accuracy with the help of Gradient boosting algorithms through which better result will be obtained comparatively with other algorithms.

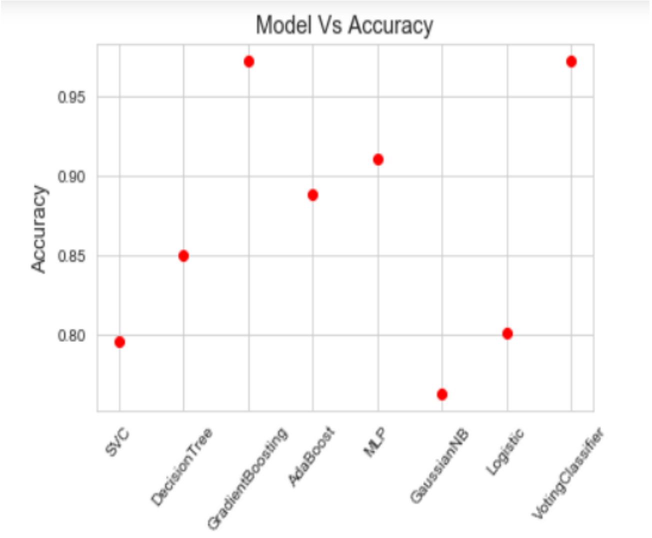


Fig.8 Comparative analysis

VI. CONCLUSION AND FUTURE ENHANCEMEN

To conclude, the analysis of attributes by which alzheimer's disease can be predicted earlier. By employing data through datasets, the correlation between attributes like CDR, eTIV and alzheimer's disease are monitored in a regular interval. Through this paper, the Gradient Boosting algorithm is discovered in order to predicts the disease with more accuracy. The advantage of Gradient Boosting alogorithm, gives more flexibility without obtaining the processed data. Furthermore, the future enhancement can be made by using clustering algorithm which may give better solution.

REFERENCES

[1] AlexandreSavio, MaiteGarcía-Sebastian, Manuel Grana, JorgeVillanua [2014],"Results of an Adaboost approach on Alzheimer's Disease detection onMRI",Research partially supported by Saiotek research projects BRAINER and S-PR07UN02, and the MEC research project DPI2006-15346-C03-03.

[2] SaruarAlam, Goo-Rak Kwon, and Chun-Su Park[2017],"Twin SVM-Based Classification of Alzheimer's Disease Using Complex Dual-Tree Wavelet Principal Coefficients and LDA",Journal of Healthcare Engineering Volume 2017.

[3] Linda Mary, John Ashima Sharma, Siddhant Gujarathi[2019],"Detector and Predictor System for Alzheimer's Disease using Naives Bayes and Decision Tree Algorithm",IOSR Journal of Engineering (IOSRJEN),ISSN (e): 2250-3021, ISSN (p): 2278-8719,Vol. 09, Issue 04.

[4] Barros, Rodrigo C, Basgalupp, Carvalho, Freitas[2012],A Survey of Evolutionary Algorithms for Decision-Tree Induction. IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, vol. 42.

[5] S.R.Bhagya Shree, H.S.Shshadri [2018], "Diagnosis of Alzheimer's disease using Naive Bayesian Classifier. Assoiations of computer machinery (ACM) on Neural Computing and Applications,Vol.29.No.1.

[6] Piers Johnson, Luke Vandewater, William Wilson, Paul Maruff, Greg Savage, Petra Graham, Lance S Macaulay, Kathryn A Ellis, Cassandra Szoeke, Christopher C Rowe, Colin L Masters, David Ames, Ping Zhang[2014],"Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease Algorithm",Johnson et al. BMC Bioinformatics,15(Suppl 16):S11.

[7] Luiz K. Ferreira, Jane. Rondina, Rodrigo Kubo Carla, R. Ono Cl,M. Rondina·Rodrigo KuboCarla,R. Ono·ClaudiaC. Leite,Jerusa Smid·Cassio Bottin,Ricardo Nitrini·Geraldo[2018],"Support vector machine-based classification of neuro images in Alzheimer's disease: direct comparison of FDG-PET, rCBF-SPECT and MRI data acquired from the sameindividuals",Rev.Psiquiatr. vol.40 n2.

[8] S.R.BhagyaShree, H.SSeshadri, [2019],"Diagnosis of Alzherimer's Disease using Naive Bayesian Classifier", Neural Comput&Applic DOI 10.1007/s00521-061-2416-3.

[9] S.Naganandhini P.Shanmuga vadivu, [2019] "Effective Diagnosis of Alzheimer's Disease using Modified Decision Tree Classifier", Volume 165.

[10] Dr.ManasiPatil, Anil Yardi[2011]," MLP classifier for Dementia Levels", DOI: 10.7763/IJMO.2012.V2.184