Dementia Prediction on OASIS Dataset using Supervised and Ensemble Learning Techniques

Shanmuga Skandh Vinayak E, Shahina A, Nayeemulla Khan A

Abstract—The Magnetic Resonance Imaging (MRI) data, which are a prevalent source of insight in understanding the inner functioning of the human body is one of the most preliminarymechanisms in the analysis of the human brain, including and not limited to detecting the presence of dementia. In this article, 7 machine learning models are proposed in the analysis and detection of dementiain the subjects ofOpen Access Series of Imaging Studies(OASIS) Brains 1, using OASIS 2 MRI and demographic data. The article also compares the performances of the machine learning models in terms of accuracy and prediction duration. The proposed model, eXtreme Gradient Boosting (XGB) algorithm performs with the highest accuracy of 97.87% and the fastest prediction durationof 0.031s/sample.

Keywords—Dementia, detection, Machine Learning, Algorithms, OASIS, feature selection, dimension reduction.

I. INTRODUCTION

Dementia – A severe disorder that impacts the memory, thinking and communication capability of the brain,that affects over 50 million individuals world-wide according to Statista [1]. It is predicted to accrue to an approximate 152 million by the year 2050.

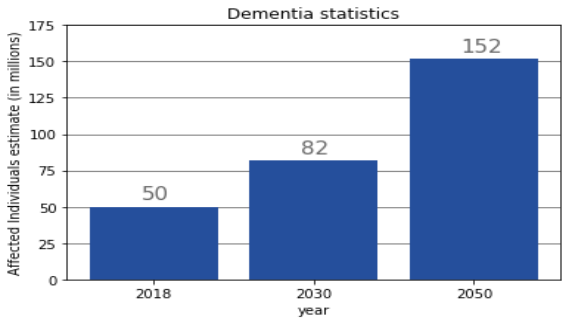

Fig. 1. Dementia disease statistics

According the World Health Organization (WHO), there are approximately 10 million new cases every year, out of which 60% - 70% of the cases result in Alzheimer's disease[2]. Although presumed to affect the elder (above 65 years old) population, 9% of the cases are accounted for the younger (below 65 years old) population as well. Dementia is caused due to damage to the nerve cells and the connections in the brain. Although dementia and the diseases caused by it are classified to be untreatable, some causes of dementia symptoms such as, infections, metabolic problems, brain tumours, anoxia, etc.are deemed curable with the appropriate treatment [3].

Revised Manuscript Received on October 10, 2020.

Shanmuga Skandh Vinayak E*, Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India. Email: shanmugaskandhvinayak16095@it.ssn.edu.in

Shahina A, Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India. Email: shahinaa@ssn.edu.in

Nayeemulla Khan A, School of Computing Sciences and Engineering, Vellore Institute of Technology, Chennai, India.Email: nayeemulla.khan@vit.ac.in

Hence, a fast and simple system capable of identifying the presence of dementia, utilizing the clinical and demographic data of a person could be effective in providing swift diagnosis. In this experiment, the Magnetic Resonance Imaging (MRI) scans data along with demographic assessments data such as the Mini-Mental State Exam (MMSE) scores, the education level of the subject, the socio-economic status of the subject, etc. collectively named as MR(Magnetic Resonance) session, are considered in developing machine learning models that predicts the presence of dementia in the subject.

This paper is organized as follows. Section II reviews the contemporary studies carried out on the detection of dementia using machine learning techniques. The Experimental setup and results are discussed in section III. The limitation of this work along with the further directions are discussed in the concluding section IV.

II. RELATED WORKS ON DEMENTIA DETECTION

Deepika Bansal et al., in their work "Comparative Analysis of Various Machine Learning Algorithms for Detecting Dementia" [9], provide comparative analysis on the algorithms utilized, to predict dementia for the OASIS 1 dataset from the OASIS 2 trained prediction models. The authors propose 4 machine learning algorithm models (J48, Naïve Bayes, Random Forest and Multilayer Perceptron), with selected features to predict dementia. Out of the 4 models, the J48 algorithm predicts the classifications with an accuracy of 99.52% for OASIS 1 dataset and an accuracy of 98.66% for the OASIS 2 dataset. The authors Yudong Zhanget al., in their work "Classification of Alzheimer Disease Based on Structural Magnetic Resonance Imaging by Kernel Support Vector Machine Decision Tree" [10], illustrate the working of an SVM based decision tree (DT) model, that is capable of detecting the presence of Alzheimer's disease (AD) in the OASIS 1 – Cross sectional data. The metadata obtained from the 3-D image source of OASIS dataset is utilized in training the model. The model combines the working of an SVM algorithm with the decision tree algorithm,to separate classification classes at each level of the decision tree. With a 5-fold cross validation, the kSVM – DT exhibits an accuracy of 80%and a classification computation speed of 0.22s for a test subject. In the work "Prediction of Alzheimer's Disease using Oasis Dataset" by Chandni Naidu et al. [11], the authors predict the presence of Alzheimer's disease in subjects based on the CDRin the OASIS 1 dataset. 4 machine learning models (Random Forest, LASSO, Gradient Boosting, SVM) are proposed, of which the Gradient Boosting and the Random Forest algorithms provide the maximum accuracy of 97.94% in the prediction process.

III. EXPERIMENT AND RESULTS

The experiment models are trained and tested using the data obtained from the MRI scans and the demographic data obtained during each scan from the test subjects(MR session). The experiment is conducted to classify the subjects as one of the three classification groups (dementia positive, converted to a dementia state and dementia negative), using the available data. This experimentalsoaims to comparethe machine learning models based on theiraccuracy and performance to classify the subject data.

A. Dataset

The data utilized in the training process of the models are obtained from the Open Access Series of Imaging Studies (OASIS) Brains project. The OASIS Brains project provides macroscopic (whole brain) structural MRI neuroimagingdemographic datasets, open to the scientific community. These data are obtained from the MRI scans, diagnostic tests and the demographic data, collected from subjects, during thetestingprocess of Alzheimer's disease diagnosis. Two sets of data are utilizedfrom the OASIS Brains project.

i. OASIS 1 Cross-sectional MRI Data.
ii. OASIS 2 Longitudinal MRI Data.

The OASIS 1 Cross-sectional MRI dataset [4] released by Marcus et al. in 2007 [5], consists of a cross-sectional short-term collection of MR session data for a subject size of 416,consisting of young, middle-aged, non-demented and demented older adults between the ages of 18 and 96. For each subject, 3 or 4 individual T1-weighted (recovery of magnetization before measuring the MR signal by changing the repetition time (TR)) (figure 2) MRI scans are obtained, with a total of 434 MR sessions. The subjects are right-handed with100 test subjects over the age of 60 have been clinically diagnosed with very mild to moderate Alzheimer's disease (AD), indicating the presence of dementia. This is the dataset for which the classification will be predicted using the trained models.

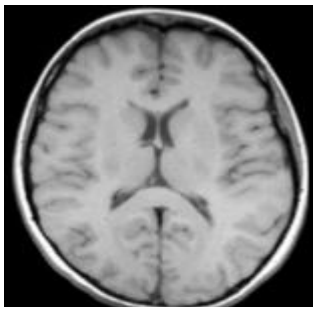

Fig. 2. T1 – weighted MRI scan

The OASIS 1 Cross-sectional MRI dataset consists of 12 features that were recorded at every MR scanning session.

TABLE I. OASIS 1 DATA FEATURE DESCRIPTION

Feature	Description
MRI ID	The identification code for each MR session.
Age	The age of the subject at the time of the MR session.
M/F	The sex of the subject.
Hand	The subject's significant hand of use.
EDUC	The educational level of the subject. The education is a 5-level categorization. i. Level 1 – subject received lower education than high school. ii. Level 2 – subject graduated high school. iii. Level 3 – subject received college level education. iv. Level 4 – subject graduated college. v. Level 5 – subject received education beyond college.
SES	The Socio-Economic Status (SES) of the subject. The status is a 5-level categorization. i. Level 1 – subject belongs to the lower class of society. ii. Level 2 – subject belongs to the lower-middle class of society. iii. Level 3 –subject belongs to the middle class of society. iv. Level 4 – subject belongs to the middle-upper class of society. v. Level 5 – subject belongs to the upper class of society.
MMSE	The Mini – Mental State Examination (MMSE) score. Range = 0 – 30, where, 0 – more likely to be demented and 30 – least likely to be demented.
CDR	The Clinical Diagnosis Rating (CDR) given to the subject by the medical professional, after initial assessments (MMSE, MRI scans). The CDR is a 4-level categorization. i. CDR = 0. The subject is cognitively normal. ii. CDR = 0.5. The subject has very mild dementia. iii. CDR = 1. The subject has mild dementia. iv. CDR = 2. The subject has moderate dementia.
eTIV	The estimated Total Intercranial Volume (eTIV) of the brain in mm^3.
nWBV	Normalized Whole Brain Volume (nWBV) in mg.
ASF	Atlas Scaling Factor (ASF). i.e. the determinant of an affine transformation matrix of the brain MRI data points.
Delay	The interval between the previous and the current MR session in days.

The OASIS 2 Longitudinal MRI dataset [6] released by Marcus et al. in 2007 [7] consist of long-term collection of MR session datafor a subject size of 150 consisting of non-demented and demented older adults between the ages of 60 and 96. The subjects were scanned on two or more visits, were separated by at least one year with total of 373 MR sessions. For each subject, 3 or 4 individual T1-weighted MRI scans are obtained in single scan sessions. The subjects are all right-handed that consists of both men and women. 72 of the subjects were characterized as non-demented throughout the MR sessions. 64 of the included subjects were characterized as demented at the time of their initial visits and remained the same for subsequent scans. 51 individuals with mild to moderate Alzheimer's disease. Another 14 subjects were characterized as non-demented at the time of their initial visit and were subsequently characterized as demented in the later visits. The OASIS 2 dataset, that has labelleddementia group classification will be utilized in training the model that can be used to predict the OASIS 1 classification.
The OASIS 2 Longitudinal MRI dataset consists of 15 features that were recorded at every MR scanning session.

TABLE II. OASIS 2 DATA FEATURE DESCRIPTION

Feature	Description
MRI ID	The identification code for each MR session.
Subject ID	The unique identification code of the subject.
Age	The age of the subject at the time of the MR session.
M/F	The sex of the subject.
Hand	The subject's significant hand of use.
EDUC	Number of years the subject has received an education.
SES	The Socio-Economic Status of the subject. The status is a 5-level categorization. i. Level 1 – subject belongs to the lower class of society. ii. Level 2 – subject belongs to the lower-middle class of society. iii. Level 3 – subject belongs to the middle class of society. iv. Level 4 – subject belongs to the middle-upper class of society. v. Level 5 – subject belongs to the upper class of society.
MMSE	The Mini – Mental State Examination score. Range = 0 – 30, where, 0 – more likely to be demented and 30 – least likely to be demented.
CDR	The Clinical Diagnosis Rating given to the subject after initial assessments (MMSE, MRI scans). The CDR is a 4-level categorization. i. CDR = 0. The subject is cognitively normal. ii. CDR = 0.5. The subject has very mild dementia. iii. CDR = 1. The subject has mild dementia. iv. CDR = 2. The subject has moderate dementia.
eTIV	The estimated Total Intercranial Volume of the brain in mm^3.
nWBV	Normalized Whole Brain Volume in mg.
ASF	Atlas Scaling Factor. i.e. the determinant of an affine transformation matrix of the brain MRI data points.
Delay	The interval between the previous and the current MR session (days).
Visit	The ordinal number of the visit, for the MR session to the testing facility.
Group	The dementia group to which the subject belongs. The group is a 3-level categorization. i. Dementia – The subject has significant dementia. ii. Converted – The subject was converted to a significant dementia state after the initial assessment. iii. Non-Demented – The subject does not have dementia.

B. Dataset Analysis

In this section, the obtained data areanalysed on the characteristic aspects, by studying the inter-relation amongst the features and the distribution of the data.

1. *OASIS 1*

Tables 3-5 describes the OASIS 1 MR session dataset distribution, classified based on the CDRfeature.

TABLE III. OASIS 1 SUBJECT CHARACTERISTICDATA

CDR	Age (years)	Sex M/F	Hand L/R
0	44.38± 24.15	120/205	0/325
0.5	71.03± 16.20	38/44	0/82
1	76.55±12.69	9/18	0/27
2	82±5.65	1/1	0/2

TABLE IV. OASIS 1 SUBJECT DEMOGRAPHIC DATA

CDR	Education (years)	SES 1/2/3/4/5	MMSE
0	3.45 ±1.22	74/103/83/61/ 4	28.97 ±1.211
0.5	2.890 ±1.266	13/24/22/22/1	25.36 ±3.55
1	2.481 ±1.34	5/6/6/8/2	21.22 ±4
2	2.0 ±1.414	0/0/1/1/0	15 ±0

TABLE V. OASIS 1 SUBJECT CLINICAL DATA

CDR	eTIV (mm^3)	nWBV (mg)	ASF
0	1480.09±156.27	0.812±0.049	1.201 ±0.128
0.5	1493.10±180.06	0.739±0.042	1.192 ±0.140
1	1490.92±123.84	0.7097±0.03	1.184 ±0.095
2	1456.5±78.48	0.684±0.027	1.20 ±0.065

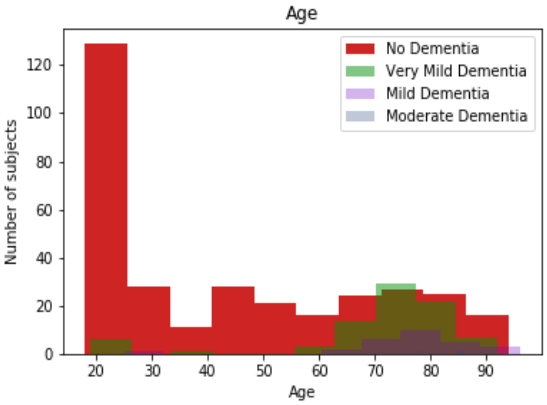

Fig. 3. OASIS 1 age distribution

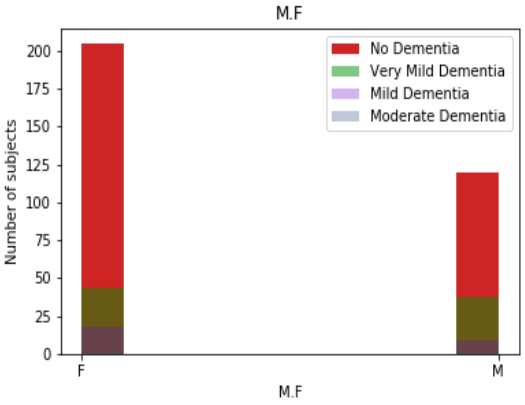

Fig. 4. OASIS 1 subject sex distribution

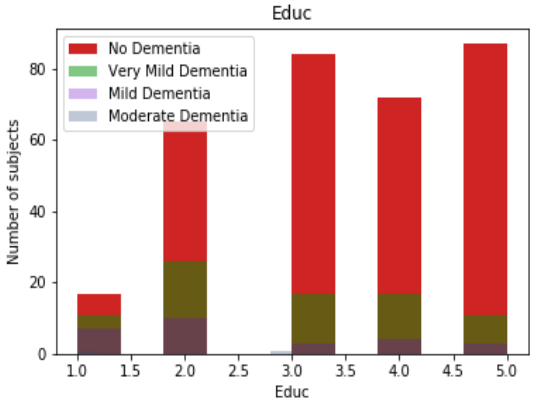

Fig. 5. OASIS 1 education level distribution

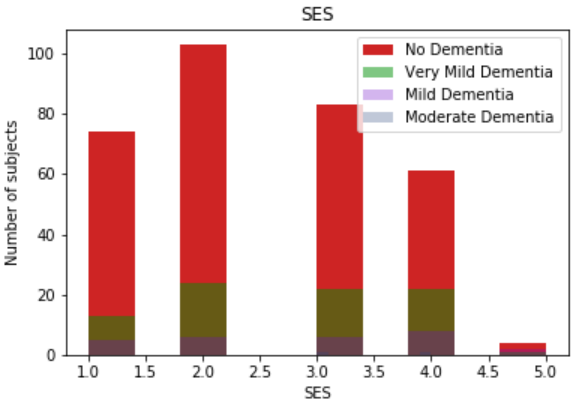

Fig. 6. OASIS 1 Socio-Economic Status level distribution

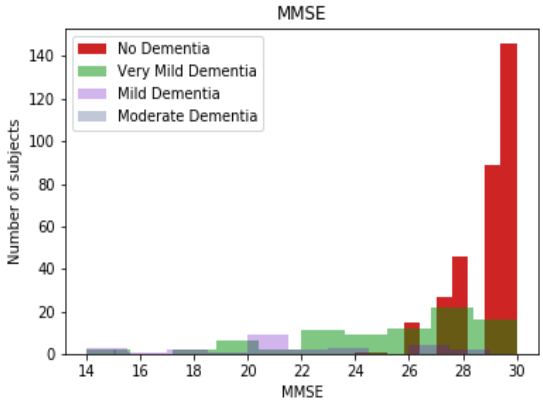

Fig. 7. OASIS 1 Mini-Mental State score distribution

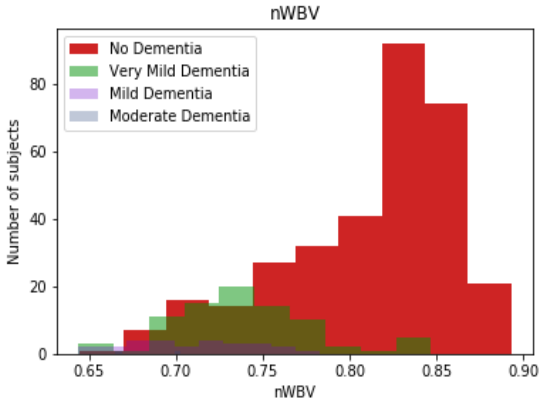

Fig. 9. OASIS 1 Normalized Whole Brain Volume distribution

From the above figures it is observed that the data for the features, age, sex, education level, SES, MMSE score and the nWBV of the subjects,have distinct and separable distribution, which influences the particular class of dementia rating. These features are further analysed in the feature selection section to obtain the optimal features to train the model.

2. *OASIS 2*

Table 6-8describes the OASIS 2 Longitudinal MRI dataset distribution, classified based on the dementia classification group.

TABLE VI. OASIS 2 SUBJECT CHARACTERISTICDATA

Group	Subjects	Age (years)	Sex M/F	Hand L/R
Demented	143	76.09 ±7.03	83/60	0/143
Converted	41	79.43 ±7.18	16/25	0/41
Non-demented	189	76.80 ±7.84	61/128	0/189

TABLE VII. OASIS 2 SUBJECT DEMOGRAPHIC DATA

Group	Education (years)	SES 1/2/3/4/5	MMSE
Demented	13.6±2.8	25/30/27 /45/6	24.3±4.58
Converted	15.51±2.56	24/9/6/2/ 0	28.41±2.41
Non-demented	15.22±2.71	40/72/43 /33/2	29.22±0.93

TABLE VIII. OASIS 2 SUBJECT CLINICAL DATA

Group	CDR	eTIV (mm^3)	nWBV (mg)	ASF
Demented	0/99/41/3	1477.81 ±163.89	0.717 ±0.032	1.202 ±0.132
Converted	18/21/2/0	1476.36 ±158.28	0.723 ±0.03	1.2 ±0.121
Non-Demented	188/1/0/0	1497.51 ±186.01	0.741 ±0.04	1.190 ±0.14

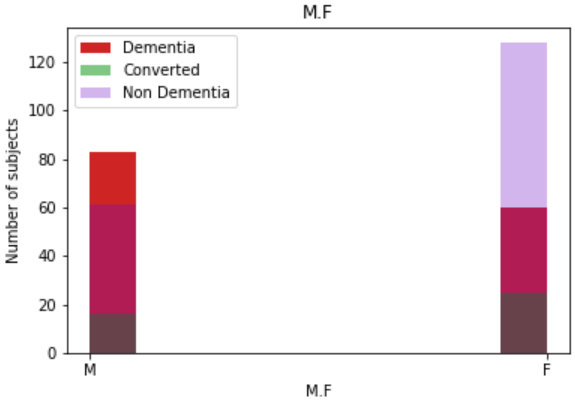

Fig. 13. OASIS 2 subject sex distribution

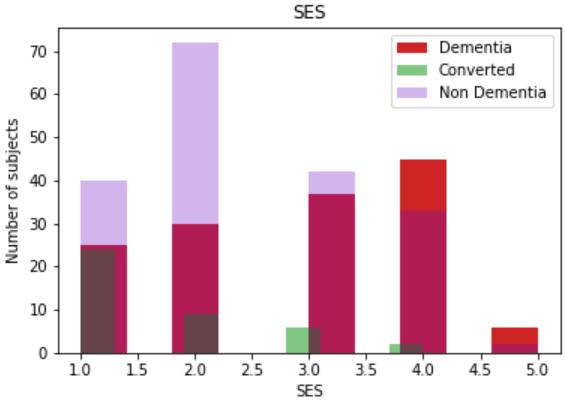

Fig. 15. OASIS 2 Socio-Economic Status level distribution

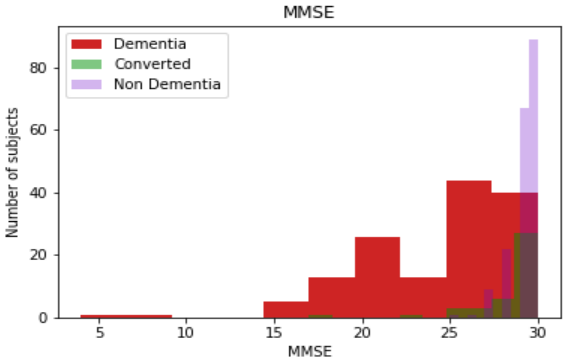

Fig. 16. OASIS 2 Mini-Mental State Examination score distribution

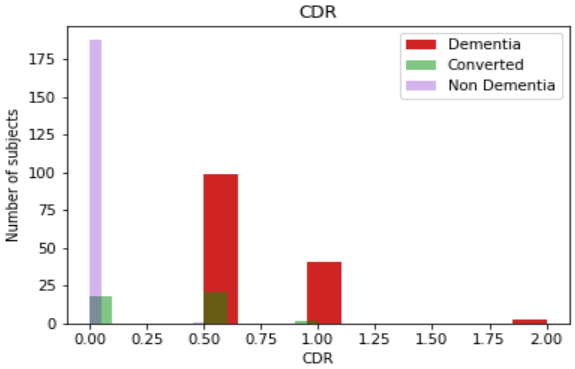

Fig. 17. OASIS 2 Clinical Dementia Rating distribution

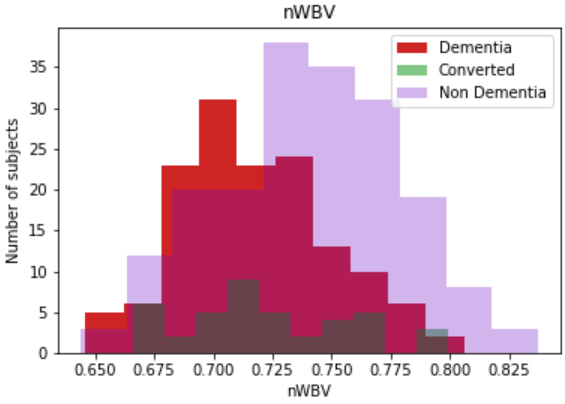

Fig. 19. OASIS 2 Normalized Whole Brain Volume distribution

From the above figures it is observed that the datafor the features, sex, SES, CDR, MMSE score andnWBV, have distinct and separable distribution, which influences the particular dementia classification group. These features are further analysed in the feature selection section to obtain the optimal features to train the model.

C. Data Preparation

The datasets, OASIS 1 and OASIS 2 possess missing data, that could affect the performance and accuracy of the training model.

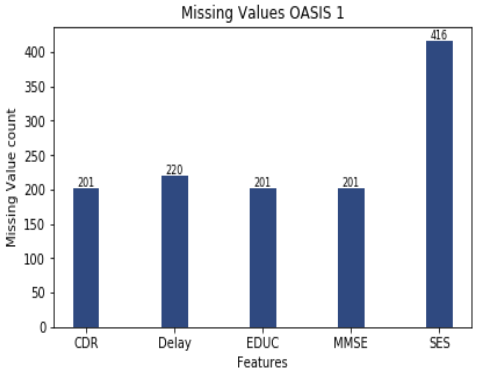

Fig. 23. OASIS 1 missing values distribution

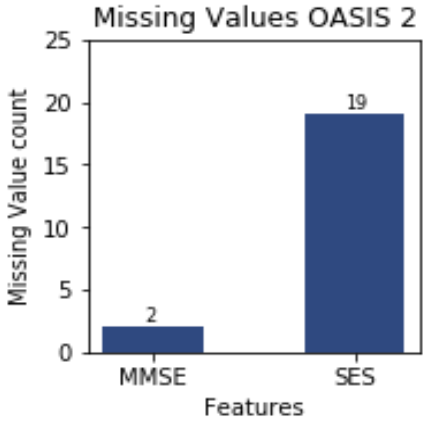

Fig. 24. OASIS 2 missing values distribution

The Multivariate Imputation via Chained Equations (MICE) package [8] in the R environment, is utilized to impute the missing values through regression modelling. The MICE tool is configured to create 10 regression models, which are then combined by obtaining the average predicted values,to impute the missing data. In the datasets OASIS 1 and OASIS 2, the categorical features,sex and hand are factorized, to be utilized by the models that does not support categories such as, Neural Network and Support Vector Machine (SVM). The non-categorical features of the datasets are normalized, to optimize the training and the prediction process of the statistical models. Equation 1 shows the normalization process for the non-categorical feature values.

$$X_{normalized} = X - X_{min} / X - X_{max} \quad (1)$$

Where, X_{min} is the feature minimum,
X_{max} is the feature maximum.

The education feature in the OASIS 1 datasetis categorized from 1 to 5, whereas in OASIS 2 dataset, the feature is the years of education. Hence, the feature in the OASIS 2 dataset is scaled between 1 and 5 using the following formula to maintain similar distribution.

$$X_{scaled} = (5-1)\frac{X - X_{min}}{X_{max} - X_{min}} + 1 \quad (2)$$

D. Feature Selection

In this section, the features of the datasets OASIS 1 and OASIS 2 are analysed to find the most optimal features that can be utilized to efficiently train the model. Instead of utilizing all the features for training and prediction, features that provide less amount information or lower the performance of the models, are eliminated.

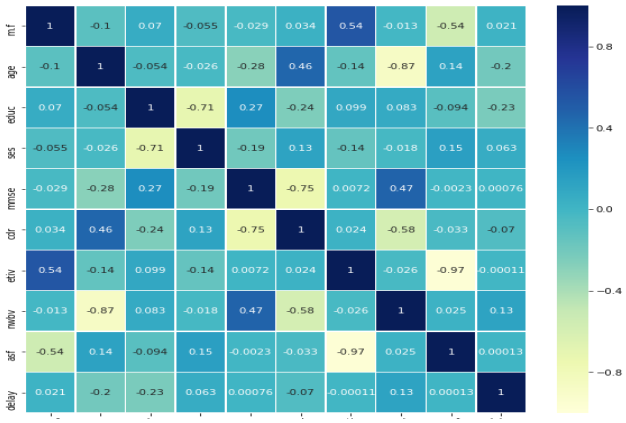

Fig. 25. OASIS 1 correlation heatmap

From the heat map it is interpreted that, theCDRexhibits the highest positive correlation with the age of the subject and also exhibits relatively high positive correlation with the SES of the subject. The CDR exhibits the highest negative correlation with the MMSE score of the subject and the nWBV of the subject and also exhibits relatively high negative correlation with the education level of the subject. Hence, these features are considered for the prediction model. From the heatmap, the feature ASFindicates to exhibit high correlation with the eTIV (figure 26). Hence, ASF is eliminated such that the model is not affected by multi-collinearity.

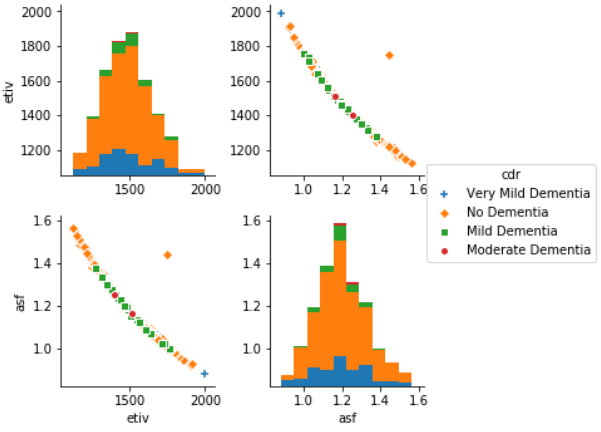

Fig. 26. High inter-correlation between eTIV and ASF of OASIS 1

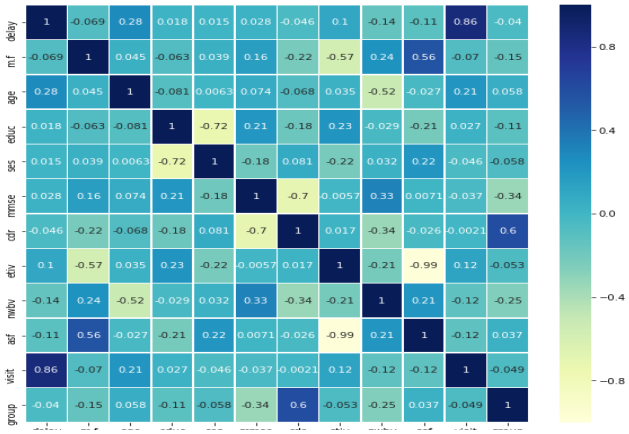

Fig. 27. OASIS 2 correlation heatmap

From the heat map it is interpreted that, the dementia classification group exhibits the highest positive correlation with CDR and also exhibits relatively high positive correlation with the age of the subject. The dementia classification group exhibits the highest negative correlation with MMSE score of the subject and also exhibits relatively high negative correlation with the sex, education level, and the nWBV of the subject. Since, the CDR correlation is also observed in OASIS 2, selecting these features is accordance with the dementia prediction for OASIS 1.

Similar to the OASIS 1dataset, the ASF feature indicates high correlation with eTIV (figure 28) and is eliminated.

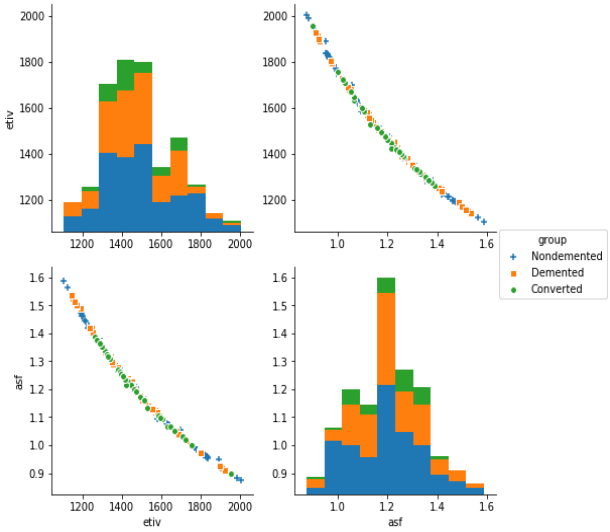

Fig. 28. High inter-correlation between eTIV and ASF of OASIS 2

The datasets are now analysed statistically to explore the best possible features for the model,utilizing the SelectKBest in scikit-learn package. The following are the final features, considered as input for the models.

i. Sex.
ii. Age.
iii. Education level.
iv. Mini-mental state examination score.
v. Clinical dementia rating.
vi. Normalized whole brain volume.
vii. Estimated total intercranial volume.

E. Dimensionality Reduction

The dimensions of the dataset are reduced, while preserving the properties that influence the classification group, to optimize the training and prediction duration. Principal component analysis (PCA) is used to reduce the dimensionsof the OASIS 1 and OASIS 2 datasets. PCA reduces the number of features, by transforming the datasets to features,that exhibit high variance, independency and orthogonality to each other.PCA of the datasets are calculated as follows

The covariance matrix of the normalized features is calculated.

$$cov(f_1, f_2) = {}^{1}/_{n-1} \sum_{i=1}^{n} (f_{1i} - \bar{f_1})(f_{2i} - \bar{f_2}) \qquad (3)$$

Where,f_1 and f_2 are any two features,
$\bar{f}$ is the average of the respective feature,
n is the number of samples.
The eigen values are obtained from the feature matrix by solving the following equation.

$$\Delta(A - \lambda I) = 0 \qquad (4)$$

Where,A is the feature matrix,
Δ is the determinant of the matrix,
$\lambda$is the eigen values
$I$is the identity matrix
Now transforming the sample to new subspace

$$y = W' \times A \qquad (5)$$

Where,W' is the transpose of the matrix corresponding to N maximum eigen values. The graphs 29 and 30 show 3 PCA projections of the OASIS 1 and 2 datasets,respectively.

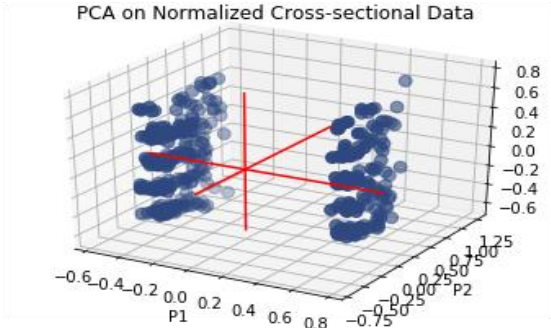

Fig. 29. OASIS 1 projections for PC 1,2 and 3

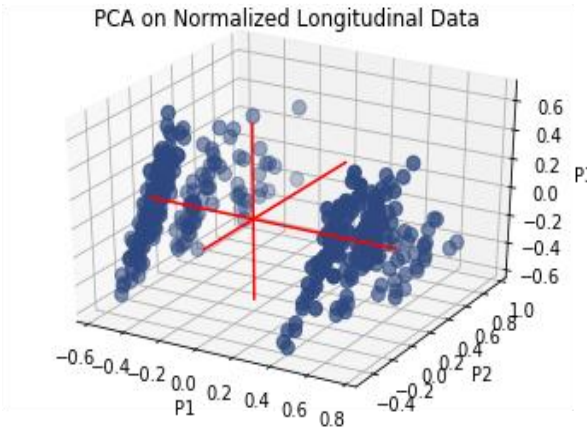

Fig. 30. OASIS 2 projections for PC 1,2 and 3

The optimal number of components are obtained by choosing a high variance knee (component 5) from the graphs 31 and 32, for the datasets.

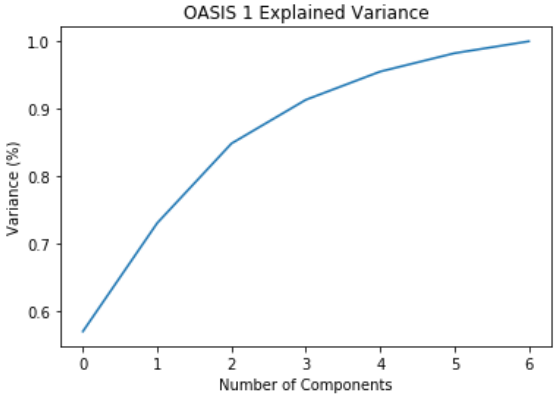

Fig. 31. OASIS 1 variance graphof principal components

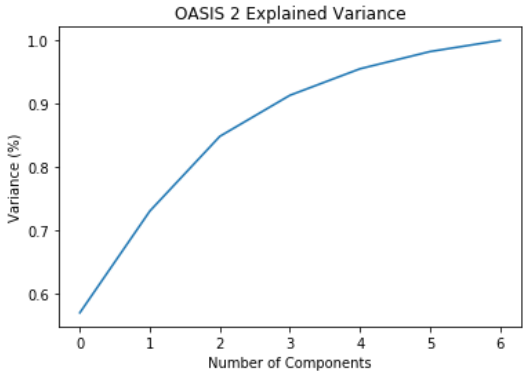

Fig. 32. OASIS 2 variance graph of principal components

F. Machine Learning Models

The dataset OASIS 2 is utilized in the dementia group training process for the following models. The dataset is split as 75% – 25%, for training and testing respectively. The models are trained and tested using 10 – k fold cross-validation.

1. *Navïe Bayes*

Naïve Bayes classifier is a probabilistic machine learning algorithm, used to classify the data based on the probability of a feature for a specific classification. Considering the 5 independent features of the dimensionally reduced OASIS dataset, the probability of the occurrence of a classification group using these features is calculated using equation 5.

$$p(Y|x_1,x_2,x_3,x_4,x_5) = \frac{p(Y|x_1)p(Y|x_2)p(Y|x_3)p(Y|x_4)p(Y|x_5)p(Y)}{p(x_1)p(x_2)p(x_3)p(x_4)p(x_5)} \quad (6)$$

Where,$p(Y|x)$ is the probability of the classification group$Y$, givenprobability of the feature x.

The maximum probability of a classification groups is selected to classify the data. The average accuracy of the Naïve Bayes model, when tested is 87.29%.

2. *k - Nearest Neighbors*

The k-NN is a supervised machine learning algorithm that classifies the data based on the clusters formed, which are labelled with the respective dementia classification group. The classification model is trained by calculating the Euclidean distances with each data points and assigning them to the cluster with the minimum distance from the cluster's centre. Euclidean distance between two data points (x and y) is calculated as follows.

$$D = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (7)$$

Where, k is the number of clusters in the model.

The optimal number of clusters is obtained by iterating through a specific cluster range and obtaining the minimum number of clusters for the maximum accuracy.

A cluster size of 19 is set and the trained model is tested to obtain an average accuracy of 90.74%.

The figure 33 shows the k-NN classification between principal components 1 and 2, where, groups 0,1 and 2 are classification groups non-dementia, dementia and converted, respectively.

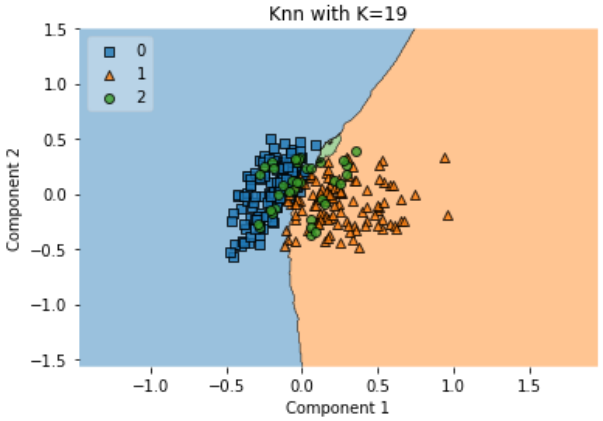

Fig. 33. OASIS 2 k-NN clusters for PC 1 and PC 2

3. *Support Vector Machine*

Support Vector Machine (SVM) is a supervised machine learning algorithm, that classifies the data using large-margin classification technique. It is a vector space-based machine learning method, where, the decision boundary between two classes having the maximum distance from any point in thetraining data, is used to classify the testing data. The SVM classifies the training data to generate the Hyperplane (decision boundaries that classifies the data points), by maximizing the distance between the data and the hyperplane. The equations8 and 9gives the hyperplane equations used in the SVM.

$$W^T X + b \geq 0 \; for d_i = +1 \quad (8)$$
$$W^T X + b < 0 \; for d_i = -1 \quad (9)$$

Where,W is a weight vector,
X is input vector,
b is bias,
$d$is the margin of separation.

The figure 34 shows the SVM separation of the principal components 0 and 1.

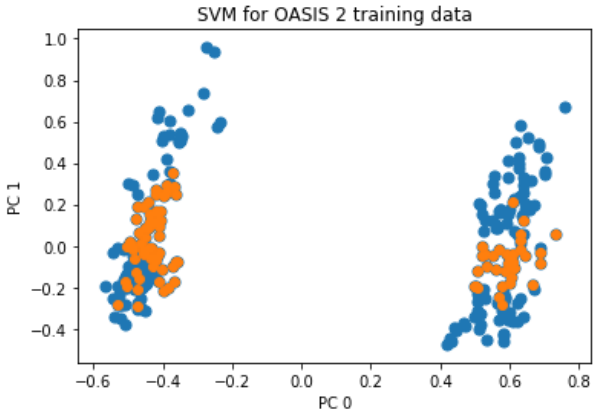

Fig. 34. OASIS 2 SVM separation graphof PC 1 and PC 2

The SVM model is trained with a linear kernel and γ = 0.001, 0.0001, which produces an average accuracy of 92.57%, when tested.

4. *Artificial Neural Network*

The artificial neural network (ANN) is a supervised machine learning algorithm that utilizes the concept of back propagation for the learning process of the data. The neural network assimilates the data distribution, to classify the same, based on the learnt dementia group classification. Each neuron in the network updates the weights and the bias of the best classification, to reproduce the states that produced the highest accuracy.

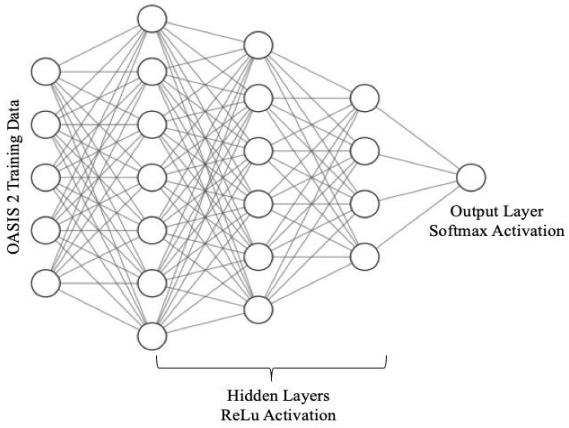

Fig. 35. Neural Network to train OASIS 2 model

The activation function of the hidden layers is rectifier linear unit (ReLu) and in the output layer, SoftMax function is used. The equations 10 and 11 show the ReLu and SoftMax equations, respectively.

$$g(z) = max(0, z) \quad (9)$$

$$S(z) = {e^{z_i}}/{\sum_j e^{z_j}} \quad (10)$$

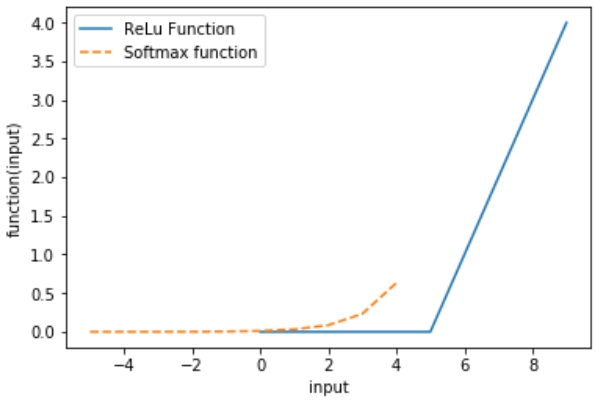

Fig. 36. ReLu and SoftMaxfunction

The ANN model is trained with the following hyper parameters.

i. Optimizer – Adam
ii. Loss function – categorical cross entropy.
iii. Epoch – 100
iv. Batch size – 10

The trained neural network model produces an average accuracy of 95.84%, when tested.

5. *Random Forest Classifier*

The random forest classifier is a supervised machine learning algorithm, that is trained using the ensemble predictions from a series of decision trees. The decision trees reach the final prediction from the features, that provide the maximum information gainat each node.

$$G(class) = E(class) - E(class, feature) \quad (11)$$

Where, $G$is the gain of information for the group,

E is the entropy(change in homogeneity) of the sample.

The figure 37 shows the root and the first level of the decision tree from the Random forest generated in the training process of OASIS 2.

Fig. 37. Decision treesof OASIS 2 training model

The trained model with 100 decision tree estimators produces an average accuracy of 96.66%.

6. *Extreme Gradient Boosting*

The eXtreme Gradient Boosting (XGB) is a decision tree ensemble technique, that parallelizes and performs greedily in the tree pruning process. Using more complex models such as,least absolute shrinkage and selection operator (LASSO) and Ridge regularization, the XBG algorithm prevents overfitting. The XGB model, utilizes the distributed weighted Quantile Sketch algorithm, to effectively find the optimal split points amongst weighted datasets. The XGB algorithm is a more robust version of the Random Forest algorithm. The trained XGB model produces an average accuracy of 97.87%.

7. *Ensemble Classification*

The Ensemble classification is a machine learning algorithm, that combines the decisions from multiple statistical models, to improve the overall performance. The ensemble model in this experimentis trained using the random forest classifier with 100 estimators and the XGB classifier. The model produces an average accuracy of 97.0%.

G. Results

The accuracies of the models that is tested with the test dataset, is given in table 9 (in descending order).

TABLE IX. MODEL ACCURACY FOR OASIS 2

Model	Accuracy(%)
XGB	**97.87**
Ensemble Algorithm	97.00
Random Forest	96.66
ANN	95.84
SVM	92.57
k-Nearest Neighbours	90.74
Naïve Bayes	87.29

The XGB model, performs with the maximum accuracy score of 97.87%. The accuracy report for the XGB model is given in the following tables and in figure 38.

TABLE X. XGB ACCURACY REPORT

Group	Precision	Recall	f1 score	Support (group count)
Non-dementia	1.00	0.98	0.99	52
Dementia	1.00	0.97	0.99	36
Converted	0.75	1.00	0.86	6

TABLE XI. TEST DATASET PREDICTED CONFUSION MATRIX

Group	Non-dementia	Dementia	Converted
Non-dementia	51	0	1
Dementia	0	35	1
Converted	0	0	6

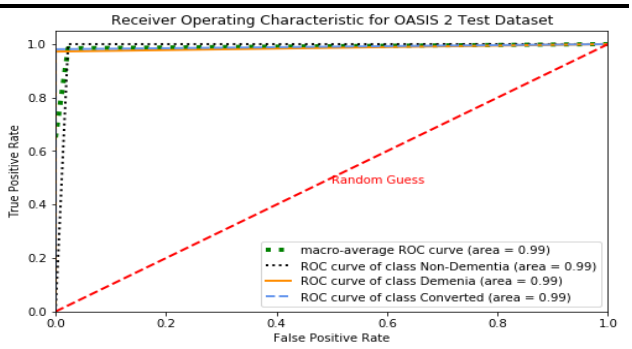

Fig. 38. Area-Under-Curve ROC of Test Dataset result

The models are also tested based on theirclassification duration of the datasets. The following table shows the durations of each algorithm to classify the test dataset of size, 94 samples.

TABLE XII. MODEL ACCURACY FOR OASIS 2

Model	Duration (seconds)
Naïve Bayes	2.21
SVM	2.58
XGB	3.0
ANN	3.30
Random Forest	3.65
XGB (without PCA)	3.72
k-Nearest Neighbours	3.88
Ensemble Algorithm	6.88

Although the XGB model is relatively slower than the Naïve Bayes and the SMV algorithm, with a processing speed of 0.31s/sample, the model is chosen to predict the OASIS 1 classification based on its high accuracy.

Utilizing the XGB model, the classification groupsof the OASIS 1 dataset is predicted. The figure 39shows the dementia classification groups in accordance to the age of the OASIS 1 subjects.

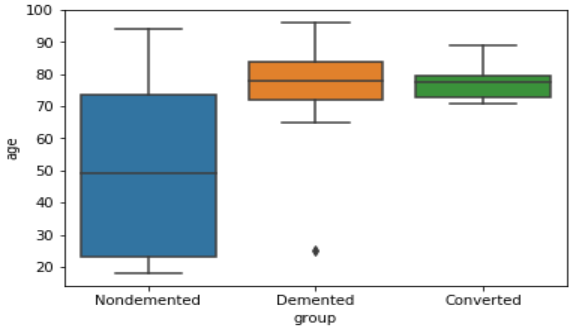

Fig. 39. Box-plot between the classification groups and the age of the OASIS 1 subjects

Since the classification group is unavailable in the OASIS 1 dataset, the CDR is taken as a deciding factor to measure the quality of the prediction model. Figure 40 shows that, the predictions are competently aligning with the CDR for each classification group, exhibiting the proficiency of the machine learning model.

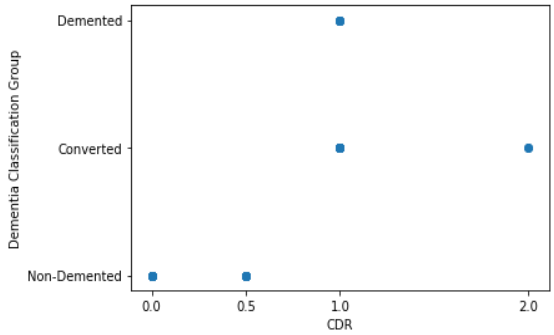

Fig. 40. Scatter plot between CDR of OASIS 1 and predicted classification groups

IV. CONCLUSION

The dementia classification group for the OASIS 1 dataset, is predicted using the XGBclassification model. The model produces the maximum testing accuracy of 97.87%, in classifying the OASIS 2 dataset, compared to the other machine learning models. The XGB model,classifies the dataset with a relatively quick time duration of 0.031s/sample, utilizing the CPU architecture to its advantage and optimallysplitting points using theQuantile Sketch algorithm.

A. Limitations:

Although this process can be utilized to aid the diagnosis process in a subject, the following factors must to be considered in the analysis process, prior to the usage of models used in this experiment and finalizing the decision.

i. The hippocampal MRI data.
ii. The prefrontal Cortex data.
iii. The family history on dementia diseases.

B. Future Works:

This study can be utilized for the OASIS 3 dataset,that consists of 19 inter-related datasets, with more than 800 features.This data can be utilized in the prediction process of a specific type of dementia such as, Alzheimer's disease.

REFERENCES

1. John Elflein, Statista, Sep 24, 2019. Accessed on: November 24, 2019. [Online]. Available: www.statista.com/statistics/264951/number-of-people-with-dementia-from-2010-to-2050
2. C. Greenblat, World Health Organization, Sep 19, 2019. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia
3. M. Clarke, M.D.&J. W. Swanson, M.D., *Mayo Foundation for Medical Education and Research*, April 19, 2019. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/dementia/symptoms-causes/syc-20352013
4. OASIS: Cross-Sectional: https://doi.org/10.1162/jocn.2007.19.9.1498
5. OASIS: Cross-Sectional: Principal Investigators: R. D. Marcus, J. Buckner& C. J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382.
6. OASIS: Longitudinal: https://doi.org/10.1162/jocn.2009.21407
7. OASIS: Longitudinal: Principal Investigators: R. D. Marcus, J. Buckner& C. J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382.
8. Stef van Buuren . Multivariate Imputation by Chained Equations. Version 3.8.0. February 21, 2020. [Online]. Available: https://github.com/stefvanbuuren/mice
9. D.Bansal, R. Chhikara, K. Khanna &P. Gupta. Comparative Analysis of Various Machine Learning Algorithms for Detecting Dementia. *Procedia Computer Science*, 132, 1497–1502, 2018. doi:10.1016/j.procs.2018.05.102
10. Y. Zhang, S. Wang &Z. Dong.Classification of Alzheimer Disease Based on Structural Magnetic Resonance Imaging by Kernel Support Vector Machine Decision Tree. *Progress In Electromagnetics Research, Vol. 144*, 171-184, 2014. doi:10.2528/PIER13121310
11. C. Naidu, D. Kumar, N. Maheswari, M. Sivagami& G. Li.Prediction of Alzheimer's Disease using Oasis Dataset.*International Journal of Recent Technology and Engineering (IJRTE)* ISSN: 2277-3878, Volume-7 Issue-6S3, April 2019.
12. P. Garrard, V. Rentoumi, B. Gesierich, B. Miller & M.L. Gorno-Tempini. Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. Cortex, 55, 122–129,2014.doi:10.1016/j.cortex.2013.05.008
13. T. Chen, A. Rangarajan &B. C. Vemuri. Caviar: Classification via aggregated regression and its application in classifying oasis brain database.*IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2010.doi:10.1109/isbi.2010.5490244
14. T. R. SivapriyaA. R. N. B. Kamal&V. Thavavel. Automated Classification of Dementia Using PSO based Least Square Support Vector Machine. *International Journal of Machine Learning and Computing*, Vol. 3, No. 2, April 2013.
15. G. Battineni, N. Chintalapudi &F. Amenta. Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM). *Informatics in Medicine Unlocked, 100200*, 2019. doi:10.1016/j.imu.2019.100200
16. B. A.Ardekani, E. Bermudez,A. M.Mubeen &A. H. Bachman. Prediction of Incipient Alzheimer's Disease Dementia in Patients with Mild Cognitive Impairment. *Journal of Alzheimer's Disease*, 55(1), 269–281, 2016. doi:10.3233/jad-160594

AUTHORS PROFILE

Shanmuga Skandh Vinayak E is a fourth-year Information Technology engineer at the SSN College of Engineering in Tamil Nadu, India. His current fields of work include image processing, signal processing and machine learning. He is interested in statistics, data science and automation.

Dr. Shahina Ais a professor in the department of Information Technology at SSN. She has 20 years of teaching and research experience. She obtained her PhD from the department of Computer Science and Engineering at IIT-Madras, India. She also has an MTech from IIT-Madras. She has research interests in the areas of Machine Learning, Deep Learning and Speech Processing. She has more than 30 research publications, including in refereed international journals and international conferences.

Dr.Nayeemulla Khan A is a Professor at the School of Computing Sciences and Engineering at VIT Chennai. He has 17 years of experience in the industry and 9 in teaching. He was the senior manager at the Airports Authority of India, when he took a break to finish his Ph.D. at IIT Madras. He then was a Research Scientist at Acusis India an MNC leading its speech recognition efforts. He has interest in the domains of Speech Recognition, Pattern Recognition and Machine Learning.