A Comparative Study for Early Diagnosis of Alzheimer's Disease Using Machine Learning Techniques

A. Bharathi Malakreddy, D. Sri Lakshmi Priya, V. Madhumitha, and Aryan Tiwari

Abstract Alzheimer's disease, a progressive neurological disorder, is one of the most common causes of dementia. This is one of the widely studied disorders to understand the changes in the brain and yet there is no cure. Having knowledge of various factors plays an important role in identifying this disease during its various stages of development. The aim of our work is to provide a system to identify the possibility of Alzheimer's disease during its early stage of progress. This paper presents the analysis of different features of the case studies, as in demented and non-demented, to derive its relation and decide the category. Later the processed data is trained on machine learning models that can fit the data well. The final model will be able to provide a well-generalized hypothesis to classify a case as either likely to be demented or not.

1 Introduction

Alzheimer's disease (AD) is a type of degenerative neurological brain disorder. It causes progressive cognitive deterioration due to deposition of beta-amyloid and neurofibrillary tangles in the cerebral cortex and subcortical gray matter [1].

Most cases of Alzheimer's disease are sporadic, entitled to the elderly with unclear etiology. Individuals with Alzheimer's disease experience noticeable symptoms like memory loss, only after years of their brain already having succumbed to the damage.

A. Bharathi Malakreddy (✉) · D. Sri Lakshmi Priya · V. Madhumitha · A. Tiwari
BMS Institute of Technology and Management, Bengaluru, Karnataka 560064, India
e-mail: bharathi_m@bmsit.in

D. Sri Lakshmi Priya
e-mail: 1by19ai014@bmsit.in

V. Madhumitha
e-mail: 1by19ai028@bmsit.in

A. Tiwari
e-mail: 1by19ai010@bmsit.in

A. E. Hassanien et al. (eds.), *International Conference on Innovative Computing and Communications*, Lecture Notes in Networks and Systems 731,
https://doi.org/10.1007/978-981-99-4071-4_16

Neurons of the brain are damaged or destroyed as the disease progresses. Ultimately, nerve cells supporting basic bodily functions, in parts of the brain, are affected and they become bed-bound.

Alzheimer's disease, being the leading cause of dementia includes symptoms like, loss of short-term memory and other cognitive deficits like, language and visuospatial dysfunction, poor judgment, and difficulty handling complex tasks due to impaired reasoning [2].

Apart from inflammation and atrophy, two of the major brain changes associated with Alzheimer's are: the accumulation of the beta-amyloid protein fragment outside neurons and abnormal form of the protein tau inside neurons [1].

Diagnosis is the most crucial and difficult part, demanding doctors with high expertise to determine dementia caused by Alzheimer's disease. Some of the diagnosis approaches include obtaining a family's medical history of cognitive, psychiatric and behavioral changes from the individual, conducting problem-solving, memory and other cognitive, physical and neurologic examinations. Brain imaging to observe the brain volume is a popular diagnosis method because brain volume shrinkage is one of the vital symptoms of Alzheimer's. Currently, cure for AD is far from possible, but its early detection can only help in ameliorating the symptoms and slow the progression of neuron damage.

2 Related Work

Chima et al. [3] suggested early diagnosis of AD using unique features like biomarkers in blood with machine learning. This approach resulted in a true positive rate > 0.79, true negative rate > 0.70 and an AUROC score > 0.80 at the initial stages of the disease.

Tarek et al. [4] used the OASIS dataset to design a convolutional neural network with six layers using Floyd hub's GPU. An accuracy of 80.25% was obtained after 545 epochs. This paper tries to address the issues with conventional machine learning algorithms that need manual feature extraction which might not be able to discern complex patterns in image data.

Alzheimer's disease cannot be diagnosed easily because the magnetic resonance imaging (MRI) data of people with Alzheimer's disease and standard healthy older people have negligible difference. Jyoti Islam et al. [5] used the OASIS dataset augmented with multiplanar patches to train a densely connected deep neural network which gave a precision of 75% for preclinical stage, 99% for non-demented stage, 62% for stage I (mild) and 33% for stage II (moderate) of Alzheimer's disease.

Some components of the brain like blood vessels and branching structures that have been affected by amyloid beta may contain pertinent information for the diagnosis of Alzheimer's disease. Conventional methods do not utilize these features. Sahrim et al. [6] proposed a method which uses branching structures of blood vessels based on tortuosity and density for the detection of AD. Computer vision techniques are used to analyze vascular abnormalities to distinguish between the features of

the tissue from people with healthy brains and those with Alzheimer's disease. An accuracy of 100% was achieved using a combination of the description of branching structures and an accuracy of 90% was achieved by using branches and their paths for classification.

Aradhana Soni et al. [7] suggest the use of a 30 s verb fluency task as a data source for diagnosis of AD. Information is extracted from the concatenated text string of verbs recorded during the task, using natural language processing. The sequence of verbs produced is used along with this information to detect AD with a recurrent neural network (RNN). An accuracy of 76% was obtained with this model.

3 System Design

We propose a method which utilizes MRI brain scan data from OASIS dataset, to detect Alzheimer's disease. The features that we have used to train the machine learning models are described in the next section.

The proposed system design has seven steps, each performing a particular task involved in building the required target model:

1. Input
2. Data visualization
3. Feature selection
4. Data transformation
5. Model training
6. Model evaluation and selection
7. Output

The workflow is presented in Fig. 1. The following subsection elaborates each of these stages in detail.

3.1 Input Dataset

The dataset was obtained from the Open Access Series of Imaging Studies (OASIS) project, aimed at studying 150 subjects who aged between 60 and 96. The study focused on longitudinal MRI data of right-handed mature individuals, with and without AD, and acquired three T1-weighted images per imaging session resulting in 373 imaging sessions, providing the imagery predictor variables. The dataset also provided non-imagery clinical predictors and demographic variables.

Considered features that represent socio-demographic attributes and clinical predictors of the subjects are listed in Table 1.

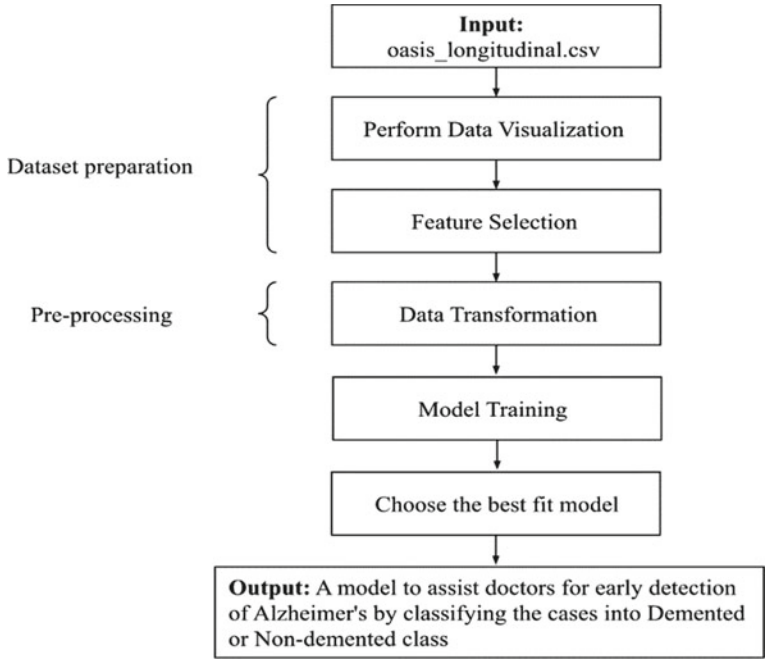

Fig. 1 Proposed system design

3.2 Data Visualization

Data visualization was performed to gain statistical and graphical insights into the data. This helped us gain a better understanding of the data, determine the structural correlation between the features, unravel outliers and inconsistencies in the structure and highlight some of the key dependencies and patterns in the data distribution.

The graphs in Fig. 2 show the amount of influence some of the attributes have on the demented and non-demented subjects. The key findings from data visualization are:

- Men are more prone to be demented than women.
- Higher strength of 70–80-years-old individuals in the demented class than that of the non-demented class.
- Non-demented group has higher brain volume when compared with demented group as evident from the graph.
- Examinations in the data presented a connection between years of education and Alzheimer's disease, indicating that demented people were less educated (in years).
- MSME graph has a higher concentration of non-demented people in the range of 26–30, whereas demented people are distributed throughout.

Table 1 Dataset description

Feature	Feature description	Operational definition
Age	Age in terms of years	[18,96]
M/F	Gender	Female—0, Male—1
EDUC (years of education)	Number of years of education completed by the subject	Primary school—1 High school—2 College—3 Undergraduate—4 Graduate and beyond—5
SES (socioeconomic status)	Socioeconomic status of the subject	Downtrodden class—1 Lower middle class—2 Middle class—3 Upper middle class—4 Upper class—5
MMSE (Mini-mental state examination)	A valid 30-point personality inventory survey that is proven to be reliable in identifying demented people	Range—[0–30] Demented—[0–24] Non-demented—[25–30]
CDR (clinical dementia rating)	This rating was acquired after a semi-structured discussion with the subject, gaining insights into their cognitive abilities	Non-demented—0 Likely to be demented—0.5 Demented—1
eTIV (estimated total intracranial volume)	This variable estimates intracranial brain volume. It is important in the study of neurodegenerative diseases because it provides a summary of variation in premorbid brain size	Observed range—[1132–1992] mm^3
nWBV (normalize whole brain volume)	The whole volume of the brain is represented here. Usually, the brain volume of people with Alzheimer's shrinks progressively and has lesser volume when compared with cognitively healthy brain	Observed range—[0.64–0.90] mg Demented range— < 0.84 mg
ASF (atlas scaling factor)	This variable is proportional to eTIV and equates to head size, allowing an estimated total intracranial volume comparison, based off, of differences in human anatomy	Observed range—[0.88–1.56]

- Demented group has higher total intracranial volume than the non-demented group.

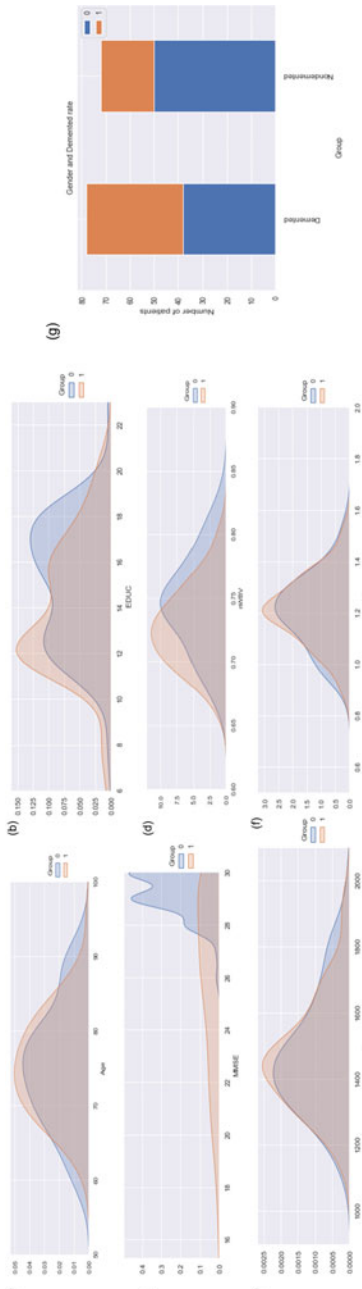

Fig. 2 **a** Distribution based on age; **b** Distribution based on years of education; **c** Distribution based on mini-mental state examination; **d** Distribution based on normalize whole brain volume; **e** Distribution based on estimated total intracranial volume; **f** Distribution based on atlas scaling factor; 0 represents non-demented, 1 represents demented in the legend of the graphs; **g** Distribution based on gender

3.3 Feature Selection

Input data obtained from the OASIS project was initially subjected to exploratory data analysis (EDA), to uncover the patterns and inconsistencies in the data distribution, which paved a way for feature extraction and selection, which is a crucial task that has a significant impact on the model's performance. After a detailed scrutiny of the data based on individual contribution and effects of correlation between the features, highly influential attributes like age, gender, years of education, socioeconomic status, brain volume ratio and MSME score were considered for further studies.

3.4 Data Transformation

Once the decision about the features was made, the data was preprocessed which involved identifying the missing data and the two approaches followed to deal with it are:

- Drop the rows with missing values
- Perform imputation- Replace the missing value with a value obtained from some chosen combining function like average or mode.
- Label Encoding - The gender column contains categorical string data which has to be numerically encoded. In this case, a simple encoding of M-1 and F-0 is done.
- Feature Scaling - Different features have different scales and ranges of input values, which when not scaled to a standard uniform range results in erroneous models. Standardization was performed on every feature so as to fit a definite scale.

Values of eight rows under the SES column were found missing. Both row dropping and imputation with median were performed to compare the performance, out of which imputation showed better results. SES is a discrete variable and median also reduces the effect of outliers, so it was chosen for imputation.

3.5 Model Training

This section deals with one of the important stages of data segregation. The ultimate goal of the project is to develop a generalized model that covers the entire population of the subset of data, providing apt results to new, unforeseen instances. For this purpose, the clean data obtained in the previous stage is split into three sets—training, validation and test set for the purpose of cross-validation. The training set is used to develop the predictive model, the validation set is used to fine-tune the model's

parameters and the test set is used to evaluate the model's performance. This ensures regularization of the model to avoid overfitting.

The models used for training the dataset are: logistic regression, SVM, decision tree, random forests, AdaBoost, averaging, max voting, bagging and boosting. A five-fold cross-validation was performed to figure out the best parameters for each model.

In the case of most neurodegenerative diseases being a life-threatening terminal disease, it is important for medical diagnostics to have a high rate of true positives for early identification of AD in patients. On the other hand, it is also equally important to make sure that the rate of false positives is as low as possible since we do not want to put the person through mental distress and the financial burden of bearing unnecessary medical therapy charges. Hence, the area under the receiver operating characteristic curve (AUC) was chosen as the main performance measure which provides an aggregate performance measure across all possible classification thresholds and displays the ability of a classifier to distinguish between two classes. The models were fine-tuned and evaluated based on its accuracy, recall and AUC scores.

Algorithms used to compare the performance of the model are listed in Table 2.

3.6 Model Evaluation and Selection

Model evaluation method is an approach to assessing the performance of each ML model. To check for the correct predictions, accuracy and metrics such as recall, F1 score, area under the ROC curve (AUC) obtained from the confusion matrix (CM) are used.

From Table 2, it is evident that random forest has the highest accuracy, recall, F1 score and AUC with 86.84%, 80%, 86.49% and 87.22%, respectively and has outperformed all other algorithms with high recall and accuracy rate, which perfectly aligns with our goal of building a model with a low number of false negatives and maintaining a good balance between precision-recall trade-off. The confusion matrix of the model trained using the random forest algorithm is shown in Fig. 3.

Hence, random forest is selected as the best classifier to build a predictive model that classifies a case as demented or non-demented.

4 Future Enhancements

Alzheimer's disease can occur due to various reasons that have not been clinically diagnosable till date. Therefore, it is important that the model is highly generalized, fitting a vast population of data. Though the accuracy of the proposed model is quite good, it can be enhanced by overcoming the present limitations of the project.

Table 2 Models trained

Model	Model architecture	Accuracy (%)	Recall (%)	F1 score (%)	AUC (%)
Logistic regression (with imputation)	Used a regularization parameter C with a value of 10 obtained after cross-validation of fivefold. This model is trained on data after filling the empty fields with the median values of the respective columns	80.56	75	81.08	81.94
Logistic regression (with dropna)	Used a regularization parameter C with value 10 obtained after cross-validation of fivefold. This model is trained on data after dropping rows with missing values	76.32	70	75.68	76.67
Support vector machine	Used RBF kernel with a regularization parameter C with the value 100 and kernel coefficient, gamma with value 0.1, obtained after cross-validation of fivefold	81.59	70	80	82.22
Decision tree	A decision tree of depth 1 was built with MSME at the root node	81.58	65	78.79	82.50
Random forest	Built a random forest classifier with 14 trees, each with a depth of 7	86.84	80	86.49	87.22
AdaBoost	Used two estimators with a learning rate of 0.0001 to provide the best model	86.84	65	78.79	82.50
Max voting (using new untrained models)	Built using three estimators with linear regression, XGB regression, random forest algorithms and arithmetic mode as a combining function	84.21	70	82.35	85
Bagging	Used four estimators built over the XGBoost algorithm which gave the highest accuracy	84.21	70	82.35	85
Averaging (using new untrained models)	Used three estimators with linear regression, XGB regression, random forest algorithms and arithmetic mean as combining functions	84.21	70	82.35	83.89
Max voting (using previously trained models)	Built using three previously trained models, decision tree, logistic regression and support vector machine with arithmetic mode as combining function	81.58	65	78.79	82.50

(continued)

Table 2 (continued)

Model	Model architecture	Accuracy (%)	Recall (%)	F1 score (%)	AUC (%)
Gradient boosting	Used five estimators Built over gradient boosting algorithm to give the best possible accuracy	76.32	65	74.29	85.56
Averaging (using previously trained models)	Built using three previously trained models, decision tree, logistic regression and support vector machine with arithmetic mean as combining function	84.21	70	82.35	85

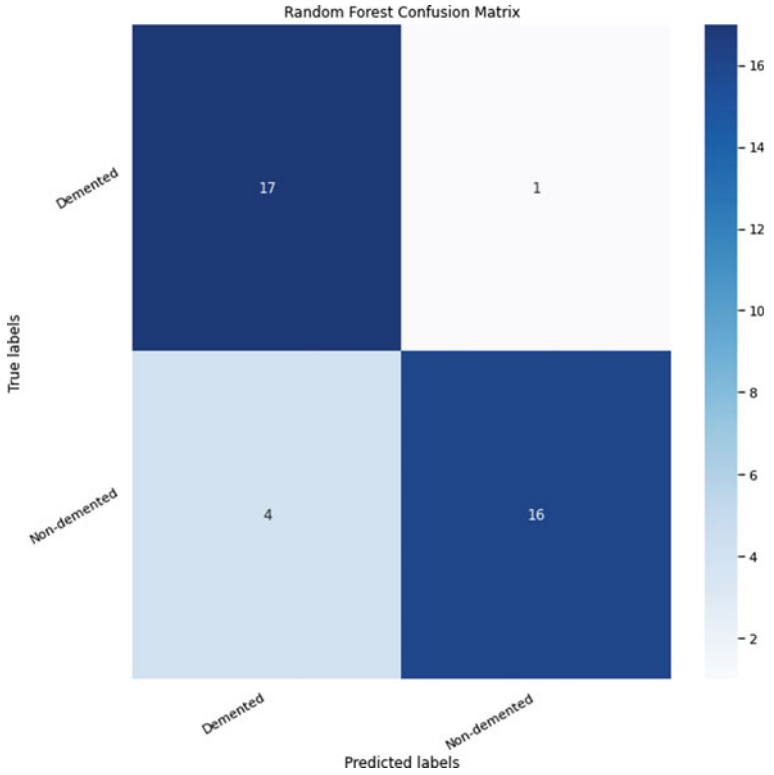

Fig. 3 Random forest model's confusion matrix

- Our study is restricted to a small population. Increasing the size of the dataset improves the predicting capability of the model by learning more patterns.
- In our model, all the features are equally weighted. Differential weighting based on the influence of each feature improves the model.

- Finding and including a broader set of relevant features also adds to models' performance.

5 Conclusion

In this project, various machine learning techniques were tested for their potential to efficiently support the prognosis of Alzheimer's disease. The proposed model serves as an accurate tool for initial screening for further medical diagnosis. The proposed framework learns the patterns of diagnosis of people at risk of Alzheimer's disease with the help of significant features imputed with mean and uses a random forest classifier that provides the highest classification accuracy of 86.84% over all other classifiers, to automate the early diagnosis of Alzheimer's disease by classifying the instances as demented or non-demented.

References

1. https://www.msdmanuals.com/en-in/professional/neurologic-disorders/delirium-and-dementia/alzheimer-disease
2. https://alz-journals.onlinelibrary.wiley.com. https://doi.org/10.1002/alz.12068
3. Eke CS, Jammeh E, Li X, Carroll C, Pearson S, Ifeachor E (2021) Early detection of Alzheimer's disease with blood plasma proteins using support vector machines. IEEE J Biomed Health Inf 25(1)
4. Ullah HMT, Onik ZA, Islam R, Nandi D (2018) Alzheimer's disease and dementia detection from 3d brain MRI data using deep convolutional neural networks. In: 2018 3rd international conference for convergence in technology (I2CT)
5. Islam J, Zhang Y (2018) Early diagnosis of Alzheimer's disease: a neuroimaging study with deep learning architectures. In: 2018 IEEE/CVF conference on computer vision and pattern recognition workshops
6. Sahrim M, Nixon MS, Carare RO, Analysing morphological patterns of blood vessels for detection of Alzheimer's disease
7. Soni A, Amrhein B, Baucum M, Paek EJ, Khojandi A (2021) Using verb fluency, natural language processing, and machine learning to detect Alzheimer's disease. In: 2021 43rd annual international conference of the IEEE engineering in medicine & biology society (EMBC) 31 Oct–4 Nov