Journal of Artificial Intelligence and Autonomous Intelligence Received 15-10-2024; Accepted 20-12-2024; Published 27-12-2024

Multistage Diagnosis of Alzheimer's Disease from Clinical Data Using 'Deep Ensemble Learning'

Manash Sarma learnermanash@gmail.com
Dept. of CSE, Ramaiah University of Applied Sciences
Bangalore - 560058, Karnataka, India

Dr. Subarna Chatterjee subarna.cs.et@msruas.ac.in
Dept. of CSE, Ramaiah University of Applied Sciences
Bangalore - 560058, Karnataka, India

Corresponding Author: Manash Sarma.

Abstract

The present research explores the early identification of Alzheimer's disease (AD) phases, encompassing Mild Cognitive Impairment (MCI), a transitional stage potentially facilitating disease prevention efforts. This investigation explores the diagnosis of Alzheimer's disease (AD) stages through the application of a multiclassification-based deep learning approach, in contrast to the predominant focus on binary classification methods for AD identification in existing research. The study utilizes the Alzheimer's Disease Neuroimaging Initiative (ADNI) clinical dataset, which encompasses over 2000 samples and exhibits an imbalanced distribution, with AD or Dementia representing the minority class. Deep Ensembled Learning is applied to the dataset with seven selected biomarkers to diagnose the disease stage through multiclassification. The ensemble approach is effective in enhancing reliability and demonstrates improved diagnostic performance for the AD stage, which belongs to the minority category. While the majority of prevalent studies utilize the Area Under the Curve (AUC) score to measure the performance of AD diagnosis using binary classification, this study employs both the F1 score and AUC score in the multiclassification of AD stages. The multiclassification for diagnosis yielded F1 scores of 88% for Cognitive Normal (CN), 86% for Mild Cognitive Impairment (MCI), and 86% for Alzheimer's Disease (AD) stage detection. The overall accuracy obtained is 87%, while the Receiver Operating Characteristic (ROC) Area Under the Curve is 91% for CN, 87% for MCI, and 91% for AD. Performance of diagnosis of AD / dementia stage that belongs to minority samples demonstrate an increase by 6% when compared with prior research [1]. The utilization of the established ADNI dataset and an ensemble approach has enhanced the reliability of the results. F1 score and AUC are effective measures, as the dataset is imbalanced. The utilization of the essential clinical biomarkers for accurate AD stage diagnosis is both efficient and effective.

Keywords: Alzheimer's disease, MCI stage, Ensemble learning, Deep learning.

Citation: Manash Sarma and Subarna Chatterjee. Multistage diagnosis of Alzheimer's disease from clinical data using 'Deep Ensemble Learning'. Journal of Artificial Intelligence and Autonomous Intelligence. Research. 2024; 1(2):9. Pages: 122-138

1. INTRODUCTION

An individual diagnosed with Alzheimer's disease (AD) develops amyloid plaque, tau, and neurofibrillary tangles in the brain. This results in a loss of connection between neurons in the brain. The hippocampus is likely the initial site of pathology. As neurons degenerate, other brain regions are subsequently affected. This progression leads to short-term memory impairment in the early stages. The initial manifestation of this condition involves a gradual decline in short-term memory function. This deterioration is subsequently accompanied by impairments in additional cognitive domains and the onset of behavioral symptoms. The progression of the disorder is typically delineated into three distinct phases: CN, MCI, and AD. "(MCI) is attractive because it represents a transitional state between normal aging and dementia"[2]. Individuals aged 65 years or older are generally susceptible to AD, with a prevalence of 50-75%. As life expectancy increases, the global incidence of AD is also rising. To date, no significant cure for AD has been established [3]. Clinical trials for AD drugs have a 99.6% failure rate [4]. In a significant development, the Food and Drug Administration granted approval for the introduction of an innovative pharmaceutical agent on June 7, 2021 for Alzheimer's disease, marking the first such approval in several decades. However, the optimism was tempered by concerns regarding uncertain efficacy, significant adverse effects, a broad target patient population, high costs, and a potential association with increased demand for diagnostics [5]. The sole viable approach remains the timely identification of the disease's progression. The primary objective of our research is to prevent deterioration through early diagnosis of AD.

ML technology is becoming a significant support in the diagnosis of diseases and their stages from biomarker datasets. The proliferation of Alzheimer's disease (AD) datasets has led to the widespread adoption of machine learning techniques in disease diagnosis. Clinical biomarker AD datasets encompass various categories, including: brain structural integrity assessments via MRI ROI; primary cognitive evaluations; cellular metabolism measurements using FDG PET ROI averages; amyloid-beta load quantification in the brain through AV45 PET ROI averages; biomarkers for brain tau load assessment; axon-related microstructural parameter evaluations employing DTI ROI; CSF biomarkers for tau and amyloid quantification; and supplementary factors such as demographic data and APOE status, including APOE4 allele count. The APOE4 allele increases the risk of late-onset Alzheimer's disease [6]. Clinical expert diagnosis from biomarkers is dependent on the focus on biomarkers from experience. This approach is time-consuming and typically prone to human error. Consequently, Machine Learning-based disease diagnosis models are gaining popularity for the diagnosis of AD and other diseases.

There are other challenges. A biomarker set generally has missing data and data may be unformatted as well. Selecting efficacious biomarkers from an extensive array of candidates poses a unique and complex challenge in the field. It is noticed that a model even trained with same dataset gives different test result. In this paper we have used feature selection techniques to automatically find only the essential biomarkers and deep ensemble learning as multiclassification technique for identifying the stage of Alzheimer's disease. Another challenge is dealing with imbalanced dataset where multiple class categories are available in the data.

1.1 Related Work

Machine learning (ML) techniques have been extensively employed in a substantial body of research focused on the diagnosis of Alzheimer's disease (AD). As this study focuses on Alzheimer's disease (AD) stage diagnosis from clinical data, we have examined previous research on clinical biomarkers while excluding genetic, epigenetic, and other pre-clinical biomarkers.

R. Chaves et al. developed a classifier for AD diagnosis employing "association rule mining" [7–10], using a SPECT dataset. They achieved an accuracy of 95.87% with 100% sensitivity and 92.86% specificity. However, the dataset was not pathologically confirmed, and the total number of samples was limited. The dataset comprised 97 samples, consisting of 43 cognitive controls and 54 AD patients. In their research, Liu, Zhang and colleagues introduced a classifier employing "Sparse representation type" (SRC) [11], aiming to develop sub-classifiers based on local patches. To improve, they applied ensemble technique on sub-classifiers. Muehlboeck et al. did research with CSF and baseline MRI together to enhance accuracy of classification [12]. The study utilized a dataset consisting of 96 samples, encompassing 273 CN patients and 96 AD patients. Employing a combination of CSF and MRI data, the proposed classification method achieved an accuracy of 91.8%. For feature extraction and identification of the ROI, Veeramuthu et al. implemented the "Fisher Discriminants ratio" technique [13]. They considered a final threshold for classification of AD and CN, had achieved 91.33% accuracy, 100% specificity, 82.67% sensitivity. But they have not discussed regarding missing data handling, handling data with imbalanced class, validation. In their research, Lawrence V. Fulton and colleagues employed a GBM algorithm to forecast the occurrence of Alzheimer's disease (AD). The model incorporated factors such as gender, age, educational background, socioeconomic status, and MMSE results. Additionally, their study utilized a ResNet-50 architecture to predict the presence and intensity of CDR through the analysis of MRI scans. While accuracy is good, MCI stage omission in investigation and generalizability in the findings are found as limitations [14]. In their research, Eufemia Lella and colleagues constructed a machine learning framework aimed at categorizing Alzheimer's disease and evaluating the significance of various features. ANN, SVM, RF techniques were applied on connectivity networks from ADNI data samples collected from AD and CN natives. They achieved AUC score 83 % and an accuracy of 75 % [15]. Sarma et al. [1], employed diverse machine learning methodologies to predict Alzheimer's disease stages CN, MCI, AD. The study demonstrated notable F1 scores of 89%, 84%, and 80% for CN, MCI, and AD stage identification, respectively, through the application of deep learning techniques on ADNI baseline clinical study data encompassing 2000 participants. To address dimensionality reduction, the investigators employed Sequential Floating Backward Selection (SFBS) methodology in conjunction with correlation matrix analysis. Nevertheless, deep learning techniques are generally susceptible to the characteristics of training data, an aspect that was not addressed in their study. These models are also susceptible to random initialization, a factor that is frequently overlooked and has not been adequately addressed. Bae et al. [16], conducted research to identify Alzheimer's Disease (AD) from T1-weighted MRI images of the medial temporal lobe. The study utilized Inception-v4, a 2D image classification CNN pretrained model. Two datasets were employed: one from ADNI and another from the Seoul National University Bundang Hospital (SNUBH). The fine-tuned model was trained using 156 AD patients and 156 cognitively normal (CN) controls from each dataset, with the final model tested on 39 AD patients and 39 CN subjects from each dataset. Five model instances were constructed from each dataset using 5-fold cross-validation. The study's conclusive findings were obtained by aggregating the mean probabilities of models developed through cross-validation. The research demonstrated an Area Under the Curve

(AUC) of 0.94 for the ADNI-trained model using ADNI test data, while the SNUBH-trained model achieved an AUC of 0.91 with SNUBH test data. When applying the ADNI-trained model to SNUBH test data and the SNUBH-trained model to ADNI test data, the resulting AUC scores were 0.88 and 0.89, respectively. Fathi et al. [17], selected six of the best individual CNN-based classifiers to combine and construct an ensemble model for classifying AD stages. The investigation yielded accuracy rates of 98.57, 94.22, 96.37, 99.83, 93.88, and 93.92 for NC/AD, EMCI/LMCI, NC/EMCI, and LMCI/AD, four-way and three-way classification groups, respectively, utilizing the ADNI MRI dataset. While the ensemble method markedly improved performance relative to individual models, its efficacy diminished when applied to a local MRI dataset. It is noteworthy that the study's reliance on accuracy as a performance metric for multiclassification deviates from current best practices in the field. Researchers have introduced G2D, a novel VLP framework that demonstrates substantial improvements in feature granularity and grounding accuracy compared to current medical VLP methods. This innovative approach exhibits exceptional performance across 25 diseases and 6 medical imaging tasks [18]. Belay et al. [19], proposed an ensemble DL model with quantum machine learning for AD classification utilizing ADNI1 and ADNI2 MRI data. They used 5 qubit quantum hardware or simulator, utilized the QSVM classification model from the Qiskit library and optimized it by adding the hyper parameters. They registered an accuracy of 99.89 and 98.37 F1-score in binary classification.

1.2 Research Gaps

Several gaps have been identified in previous research. Much of the earlier research relied on pathologically unverified data due to limited data availability. These studies utilized small datasets that generally lacked representation of the intermediate disease stage. While the performance of some earlier research results appears impressive, they were based on small, unverified datasets, thus reducing their reliability. Contemporary research has yielded noteworthy outcomes in the dichotomous categorization of AD and CN conditions. However, for stage identification including intermediate disease states such as MCI, the multiclassification results are less satisfactory. It has been observed that some researchers opted for binary classification by considering intermediate stages like MCI to be AD [20]. Although this approach may have yielded better results, this methodology overlooks a crucial dimension of scientific inquiry by failing to consider the disease's intermediate phases. Some machine learning algorithms employed are stochastic in nature, often leading to performance variance with different data, which is not addressed by proven techniques.

There is a need for a pathologically verified dataset with sufficient samples, a reliable model capable of diagnosing intermediate stages of the disease, and efficient performance analysis for imbalanced datasets, particularly when multiple class categories are present.

1.3 Key Contribution

This research work presents several significant contributions. The study utilized seven essential biomarkers from a total of 113, which were obtained from the authors' previous research. Furthermore, the study incorporates multiclassification to include the intermediate stage of Alzheimer's disease, a crucial disease stage for alerting individuals to take preventive measures. The implementation of 'Deep ensemble learning with resampling' yields a model with enhanced performance and

reliability in predicting Alzheimer's disease stages. Ensemble approach increased the performance of diagnosis of AD / Dementia stage that belongs to minority categories. Additionally, this is the first study that utilizes 'F1 score' for multiclassification in a validated and authentic dataset comprising over 2000 samples and more than 100 variables. The utilization of the 'F1 score' facilitates appropriate performance analysis of the model, given the imbalanced nature of the dataset

2. METHODS AND MATERIALS

The process model described in "The lightweight IBM Cloud Garage Method for data science" [21], is employed in this context.

2.1 Dataset Selection and Exploration

This study utilized ADNIMERGE, a clinical dataset from ADNI, which is accessible through the "Laboratory of Neuro Imaging" website. The ADNIMERGE dataset comprises baseline and periodic samples collected across several phases from 2092 patients. For the purposes of this research, the 2092 baseline samples were employed.

Initial dataset exploration was conducted with respect to age and APOE4 gene as per the authors' previous work. FIGURE 2- FIGURE 4 are derived from the author's previous work for dataset exploration. Majority of the natives belongs to 70-80 year of age group as shown in the FIGURE 2. The subsequent figures examine the influence of the APOE4 allele on Alzheimer's disease. APOE4 is broadly recognized as a significant contributor to AD [5]. In the human genome, an individual can possess up to two APOE4 genes. FIGURE 4 suggests that individuals diagnosed with Alzheimer's disease (AD) or dementia are likely to possess one or two APOE4 gene alleles. However, there are cases of individuals with dementia who do not have an APOE4 allele. Therefore, APOE4 is not the sole etiological factor for AD. FIGURE 1 shows race wise distribution of ADNI data participants where most of the participants are from white population.

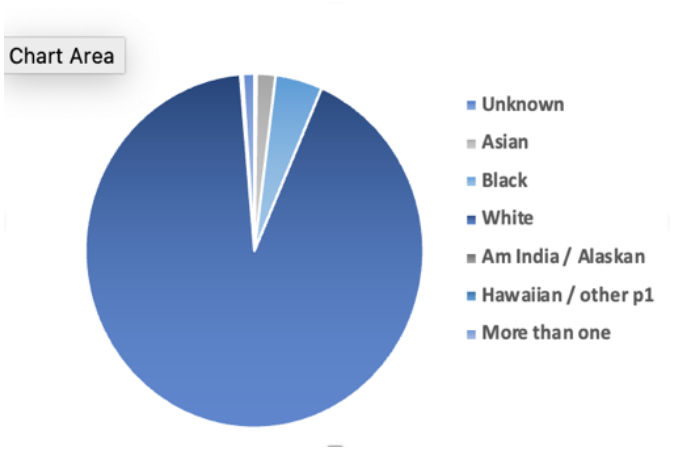

Figure 1: Race-wise distribution of ADNI participants

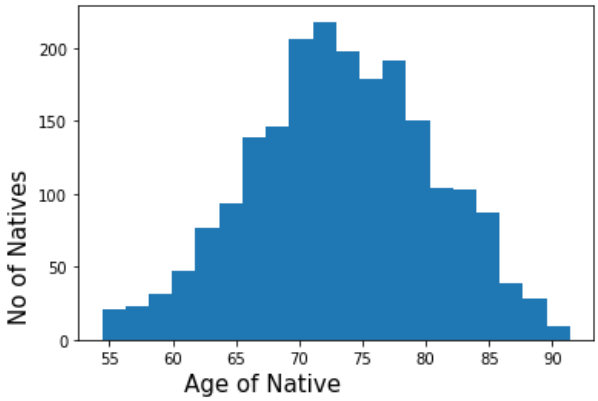

Figure 2: Age-wise distribution of ADNI participants

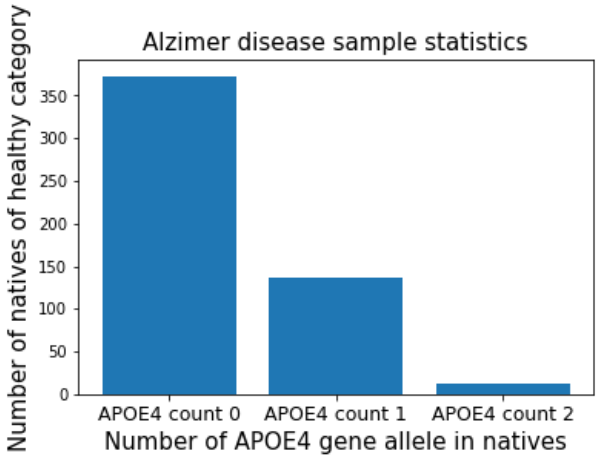

Figure 3: APOE4 count distribution in CN natives

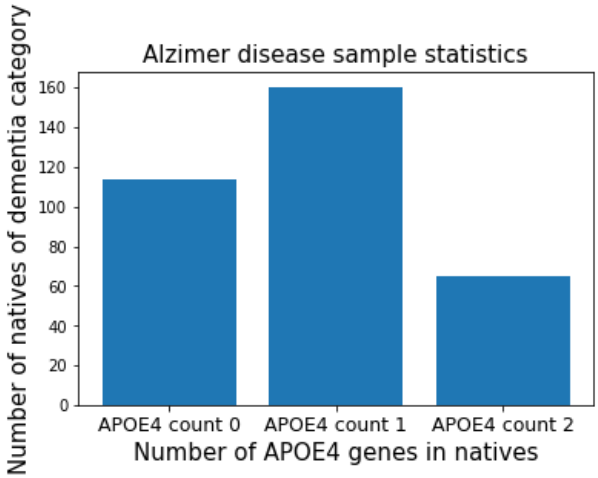

Figure 4: APOE4 count distribution in AD natives

2.2 Data Preprocessing and Feature Selection

Following the exploration of the dataset, the subsequent steps involve preprocessing, feature selection, and model construction. We employed the preprocessing and feature selection techniques outlined in the authors' previous research [1]. A brief overview of these steps is provided herein. For detailed information, please refer to the aforementioned research. The following preprocessing steps were applied to the 'ADNIMERGE' dataset, a CSV-formatted file:

- Data present in diverse types were mapped to numeric type.
- Missing values of non-target attributes were imputed with the mean value for numeric data types. This approach ensured that rows with missing values were retained.
- Rows with null target values were removed.
- Normalization of data performed subsequent to imputation and cleaning procedures.

Following preprocessing, 26 samples were eliminated from 2092 base samples. Consequently, this dataset comprises 2066 samples with the essential 7 biomarkers and the target diagnostic variable. Of these 2066 samples, 775 belong to CN, 933 belong to MCI, and 358 belong to AD or dementia category. Therefore, the dataset exhibits an imbalanced distribution of diagnostic category. Based on findings from the authors' previous research, utilizing "Sequential Floating Backward Selection" (SFBS) and "correlation matrix" methodologies, the biomarkers considered as most efficacious in Alzheimer's Disease (AD) stage classification and diagnosis are APOE4, MMSE, AGE, PTRACCAT, ADAS13, ADAS11, and mPACCdigit.

2.3 Method for Ad Stage Diagnosis Model

Over the years, Artificial Intelligence (AI) has evolved along two parallel but distinct trajectories: one concentrating on deterministic Artificial Intelligence, and the other emphasizing statistical learning approaches that have progressed into Machine Learning (ML) and Deep Learning (DL) methodologies. While deterministic AI provides clarity and precision, as every rule is explicitly defined, the lack of flexibility and dependence on pre-defined knowledge, researchers moved to ML and DL based approach with emergence of fast computational power and sufficient data. Zhai et al. [22], published a comparison of deep learning and deterministic AI algorithms. In their experiment, it was observed that the recursive randomization of the neural network weights and biases took a longer time than the direct execution of the control signal. Although, DL based models need more training time with many layers inside, availability of computational power, increasing complexity of research problems, the DL techniques are being widely accepted. Previous research has demonstrated that deep learning algorithms achieved superior performance in Alzheimer's Disease (AD) stage diagnosis compared to alternative methodologies. Deep learning models can offer flexibility, proportionally scale to the amount of training data. But traditionally, they are sensitive to the nature of the training data. They are also sensitive to the random initialization as they learn via a stochastic training algorithm. So, there may be different set of weights every time of training, resulting different predictions. Apart from high variance in the data set, there is often issue of data imbalance particularly in medical datasets. There are two popular techniques to contain data

imbalance – increasing the weightage of minority samples and oversampling the minority samples or under sampling of majority samples [23]. SMOTE is often used for oversampling or under-sampling [24]. Adjusting the weight can be complicated. In oversampling, there is a likelihood of overfitting and in under-sampling, there is risk of losing important data samples [25, 26]. Weight adjustment and sampling may not increase accuracy though it can increase recall compared with a traditional DL model. Earlier research [25], successfully utilized DEL to improve overall classification performance in imbalanced datasets and achieved better result than other ensemble learning approach. Scientists have suggested employing ensemble learning techniques that involves training multiple models and combining their predictions. An ensemble learning model's classification outcome can be generally represented by the equation below:

$$y = \max(L1(x), L2(x), L3(x), \ldots, LN(x))$$

where
max: The majority voting method used to combine the predictions from learners.
Li (x): Prediction of an individual learner Li on observation x.
N: Total number of individual learners
y: The combined prediction of the learners Li (i = 1 to N)

To mitigate the challenges of outcome variance, issue of data imbalanced, our study incorporates an ensemble learning strategy in combination with deep learning techniques. This manuscript introduces a model founded on deep ensemble learning principles. Earlier research [27], employed deep ensemble learning technique on National Alzheimer's Coordinating Center (NACC) data for AD classification and achieved better results while compared with other ensemble techniques.

2.4 Deep Ensemble Learning With Resampling

This study employed a resampling-based deep ensemble learning technique, wherein the data initially reserved for the construction and training of individual deep learning models in the ensemble was repeatedly resampled. The primary rationale for selecting a resampling-based ensemble approach was twofold:

- The training data is not divided among different models, as is the case in other ensemble approaches, thus maintaining the full extent of training data.
- The efficacy of deep learning architectures is contingent upon the availability of extensive training datasets. Given that 2092 base samples were utilized to train 25 models, dividing this data among the 25 models would significantly reduce the available training data, consequently impacting model performance.
- In this study, we deviate from conventional resampling methods like K-fold validation. Our objective is to experiment with 25 distinct model instances. Given the available 2092 base samples, employing a 25-fold validation approach would result in extremely small test sample sizes, potentially compromising the model's reliability.
- Alternative resampling-based ensemble learning method such as RF is not used here as deep ensemble learning was found to superior result in prior research.

2.5 Ensemble Model Overview With Data Organization

The processed data from the pre-processing and feature-selection step is utilized as input for ensemble model construction (and testing). As illustrated in FIGURE 5, the data is randomly partitioned into model construction data and test/evaluation data with a 75% and 25% ratio, respectively.

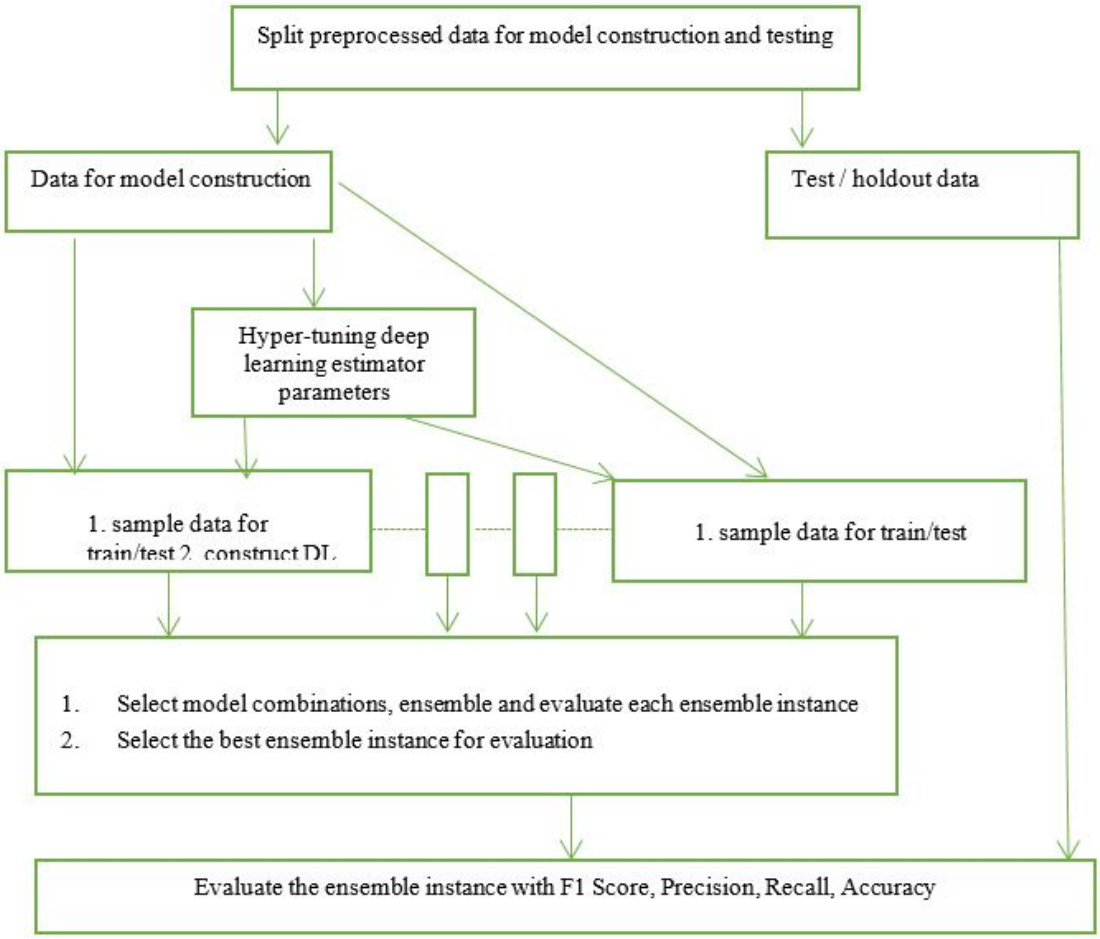

Figure 5: Model training and testing workflow

- The model construction dataset is used for training and testing each individual deep learning model where we use resampling method for repeatedly generating new set of train and test out of the model construction data from step above.
- We then combine predictions of individual models which is performance wise more stable and better than an individual model.
- After fitting and evaluating each individual model, we estimate the expected performance of the ensemble model for different combinations as mentioned in TABLE 1.
- Ensemble of sizes from 1 to 25 is evaluated from holdout dataset prepared at first step.

Table 1: Summary

Hyper tuning parameter	Parameter value
Optimizer	Adam optimizer
Cost or loss function	categorical cross_entropy
Learning rate	0.001
Batch size	5
Epochs	4000
No of layers	2
Activation function – layer 1	RELU
Activation function – layer 2	Softmax
Dropout rate	0.20

2.6 Ensemble Model Training

The ensemble framework configuration involved the independent training of 25 deep learning models. As previously mentioned, the dataset allocated for model development was partitioned into training and test subsets, constituting 75% and 25% of the data, respectively, using the 'Scikit-learn' library methodology. The high-level algorithmic procedure for model training encompassed the following steps:

i Perform resampling of training and test datasets.

ii Train and build an individual model.

iii Incorporate the model into a collective list.

iv Iterate through steps i, ii, and iii for all 25 models.

The Multi-Layer Perceptron (MLP) based Keras library is utilized for sequential model construction. FIGURE 6 delineates the training and learning configuration of an individual deep learning model (MLP) within the ensemble. FIGURE 6 illustrates the loss and accuracy of each of the four models that constitute the optimal combination of the ensemble, as elucidated in the Results section.

3. RESULT

3.1 Setup

In the ensemble framework setup, we individually train and construct a total of 25 deep learning models, which are subsequently stored in a list as described in section 2.6. For ensemble evaluation, we iteratively retrieve a model from the list and perform inference to obtain classification results using the reserved holdout data. Then we combine individual classification results of the models to calculate ensemble classification test result using maximum voted classification. TABLE 2 below lists these results. FIGURE 7 graphically shows accuracy scores of different ensembles as listed in

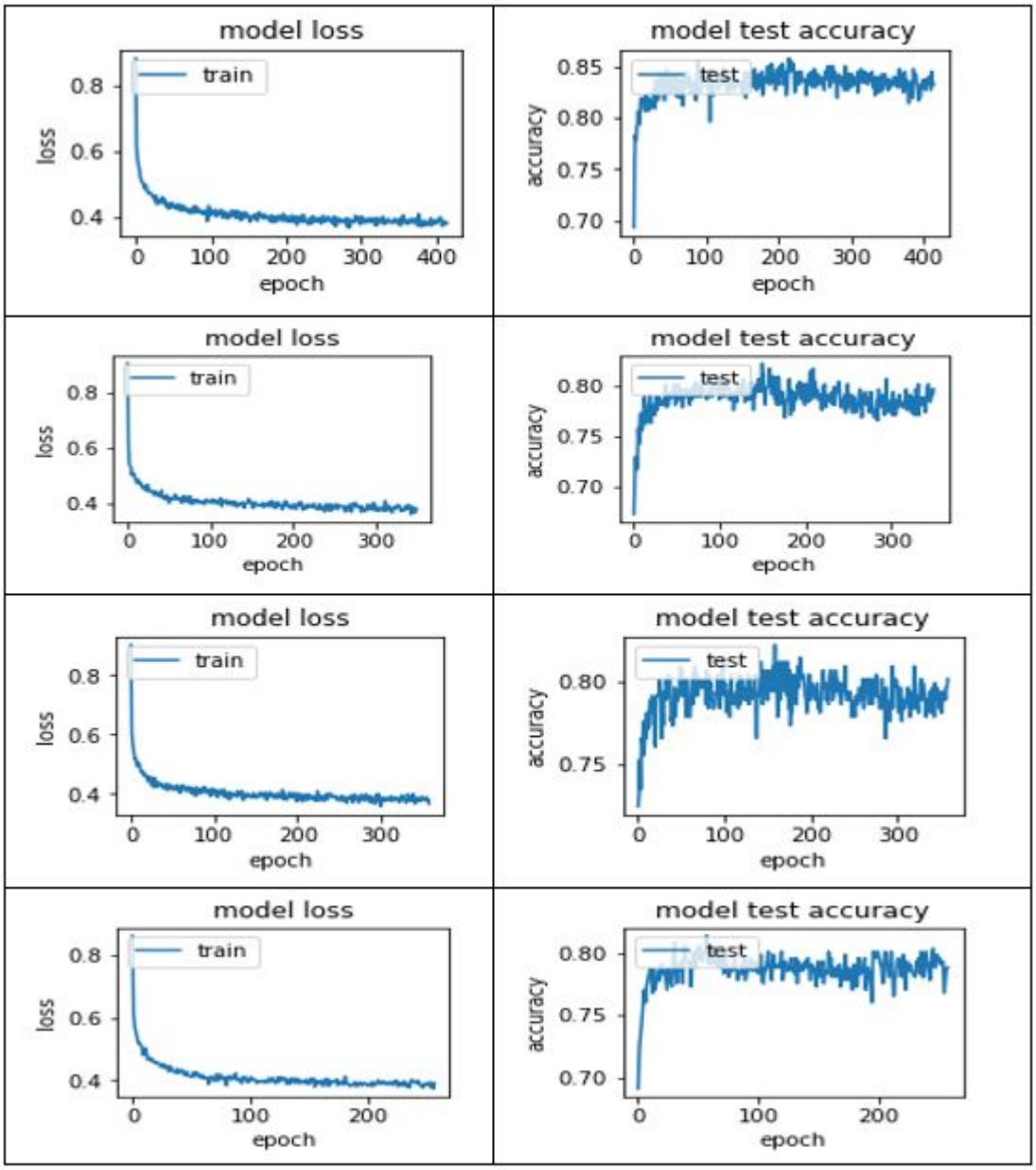

Figure 6: Loss and accuracy during training of individual models of best ensemble

TABLE 2. We achieve best ensemble accuracy score of 0.867 as shown in 4th row of table where number of models in the ensemble is 4. We established the ensemble combination of these 4 model instances to evaluate the performance of the ensemble model described in the subsequent section.

Table 2: Scores of different combinations of ensemble model

Individual model	Model score (Accuracy)	Number of models	Ensemble score
1^{st}	0.855	1	0.855
2^{nd}	0.836	2	0.851
3^{rd}	0.855	3	0.867
4^{th}	**0.865**	**4**	**0.867**
5^{th}	0.855	5	0.861
6^{th}	0.865	6	0.861
7^{th}	0.849	7	0.863
8^{th}	0.863	8	0.857
9^{th}	0.832	9	0.859
10^{th}	0.863	10	0.859
11^{th}	0.857	11	0.861
12^{th}	0.839	12	0.859
13^{th}	0.859	13	0.863
14^{th}	0.859	14	0.861
15^{th}	0.843	15	0.863
16^{th}	0.841	16	0.861
17^{th}	0.853	17	0.863
18^{th}	0.843	18	0.859
19^{th}	0.857	19	0.857
20^{th}	0.849	20	0.859
21^{st}	0.865	21	0.855
22^{nd}	0.855	22	0.855
23^{rd}	0.857	23	0.855
24^{th}	0.849	24	0.857
25^{th}	0.847	25	0.857

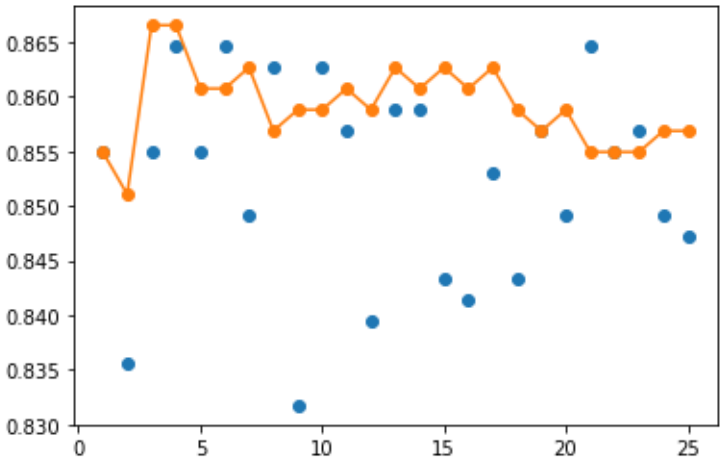

Figure 7: Ensemble model accuracy (vertical) vs number of models (horizontally) in the ensemble

4. PERFORMANCE ANALYSIS AND DISCUSSION

This study employed ROC AUC and F1 score as metrics to assess the model performance in multi-class classification. As discussed in section 2.2, the dataset exhibits an imbalance. 'Accuracy', a commonly employed performance metric, is recognized as an inadequate measure for model performance assessment with imbalanced data [28]. Accuracy is not always the most appropriate metric for performance analysis of datasets, as the accuracy score is influenced by the category with the majority population in an imbalanced dataset. Therefore, precision and recall are more suitable measures in this context. The F1 score, being the harmonic mean of precision and recall, facilitates a comprehensive performance analysis. This variant is frequently utilized as a performance metric in learning from imbalanced data. Furthermore, the ROC (Receiver Operating Characteristic) Curve and AUC (area under curve) are widely employed in performance analysis. [29, 30].

In TABLE 2, of section 3.1, the accuracy/evaluation score of an individual model ranges from 83% to 87%, while it remains within 86%-87% across all ensemble formations. This finding reaffirms our ensemble learning approach, which demonstrates the potential to produce a robust and reliable model with a better performance score. This observation can be readily discerned from the performance results of current and prior research, as illustrated in TABLE 3, TABLE 4, FIGURE 8, and FIGURE 9. For performance analysis, the best ensemble model combination, as presented in the TABLE 2, is evaluated using the hold-out data set. The evaluation employs performance metrics including F1 score, recall, and precision, as delineated in TABLE 4. ROCAUC and F1 proves useful in performance analysis while dealing balanced and with imbalanced dataset as well. In FIGURE 9, the ROC curve for each classification is close to the vertical axis that indicates good performance. The AUC values for CN, MCI, and AD categories are 0.91, 0.87, 0.91, respectively, while the F1 scores are 88, 86, and 86.

Table 3: Performance in prior work

Evaluation Method	Deep Learning
F1-Score (CN, MCI, AD)	% 89, 84, 80
Precision (CN, MCI, AD)	% 91, 85, 75
Recall (CN, MCI, AD)	% 88, 83, 85

Table 4: Performance in current work

Evaluation Method	Deep Ensemble
F1-Score (CN, MCI, AD)	% 88, 86, 86
Precision (CN, MCI, AD)	% 87, 86, 88
Recall (CN, MCI, AD)	% 89, 85, 86

These results demonstrate an improvement over a previous study in which AUC scores of 91, 85, 90 and F1 scores of 89, 84, 80 (FIGURE 8) were obtained for the same performance metrics. Our ensemble approach contributes to improving the F1 score for AD stage identification by 6%.

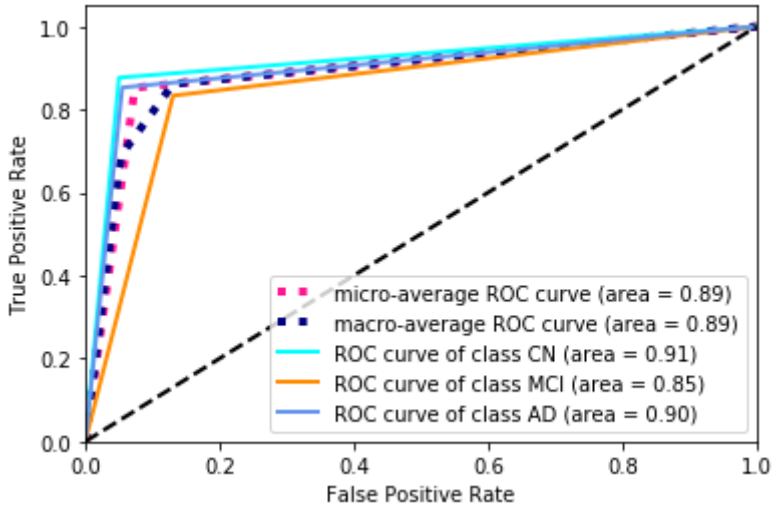

Figure 8: AUC score in prior work

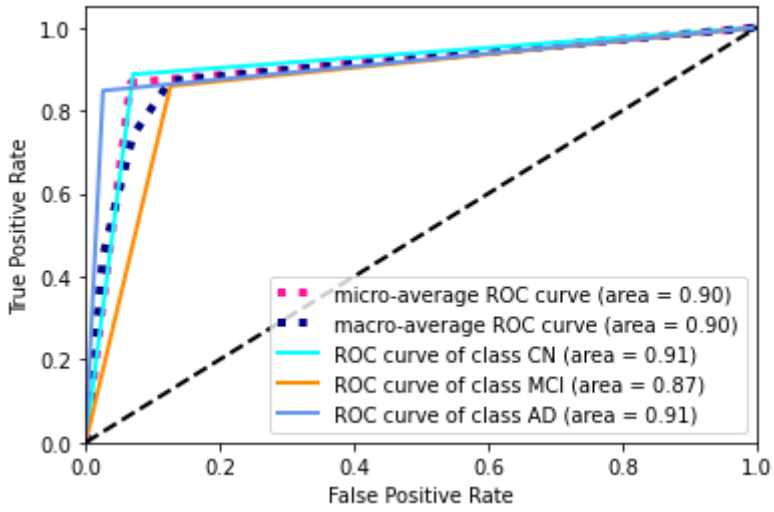

Figure 9: AUC score in current work

While an individual deep learning model can produce effective classification as described in the author's previous work, the variation of results due to the inherent stochastic nature of deep learning and dataset variability used training are addressed by the ensemble approach ensuring reliability.

To summarize, we developed a multiclassification model for diagnosing stages, including the intermediate stage of AD (MCI), based on deep ensemble learning. Several challenges have been addressed, and an F1 score was achieved that represents the highest known results thus far in multiclassification of AD stages. Utilizing the essential biomarkers from 113, appropriately applying resampling techniques with deep ensemble learning methodology enabled us to construct a reliable model with optimal multiclassification results. Furthermore, as the clinical tests for this diagnosis are associated with lower costs compared to genome sequencing and other expensive diagnostic

approaches, it is anticipated to be more accessible to low-income populations. For future research, other Alzheimer's disease datasets or potentially larger datasets could be explored. It is envisioned that this research work can provide valuable support for medical professionals in diagnosing disease stages.

References

[1] Sarma M, Chatterjee S. Identification and Prediction of Alzheimer Based on Biomarkers Using 'Machine Learning'. In Machine Learning, Image Processing, Network Security and Data Sciences: Second International Conference, MIND 2020, Silchar, India. Proceedings, Part II. Springer Singapore. 2020;1241:271-284.

[2] Angelucci F, Spalletta G, Iulio FD, Ciaramella A, Salani F, et al. Alzheimer's Disease (Ad) and Mild Cognitive Impairment (MCI) Patients are Characterized by Increased BDNF Serum Levels. Curr. Alzheimer Res. 2010;7:15-20.

[3] Crous-Bou M, Minguillón C, Gramunt N, Molinuevo JL. Alzheimer's Disease Prevention: From Risk Factors to Early Intervention. Alzheimer's res. ther. 2017;9:71.

[4] Cummings JL, Morstorf T, Zhong K. Alzheimer's Disease Drug-Development Pipeline: Few Candidates, Frequent Failures. Alzheimer's res. ther. 2014.

[5] Zissimopoulos J, Jacobson M, Chen Y, Borson S. Knowledge and Attitudes Concerning Aducanumab Among Older Americans After FDA Approval for Treatment of Alzheimer Disease. JAMA Netw. Open. 2022;5:e2148355.

[6] Saunders AM, Strittmatter WJ, Schmechel D, St. George-Hyslop PH, Pericak-Vance MA, et al. Association of Apolipoprotein E Allele □4 With Late□Onset Familial and Sporadic Alzheimer's Disease. Neurol. 1993;43:1467.

[7] Chaves R, Ramírez J, Górriz JM, López M, Salas-Gonzalez D, et al. Effective Diagnosis of Alzheimer's Disease by Means of Association Rules. In Hybrid Artificial Intelligence Systems: 5th International Conference, HAIS. San Sebastián, Spain. 2010.

[8] Chaves R, Górriz JM, Ramírez J, Illán IA, Salas-Gonzalez D, et al. Efficient Mining of Association Rules for the Early Diagnosis of Alzheimer's Disease. Phys. Med. Biol. 2011;56:6047-6063.

[9] Chaves R, Ramírez J, Górriz JM, Puntonet CG. Association Rule-Based Feature Selection Method for Alzheimer's Disease Diagnosis. Expert Syst. Appl. 2012;39:11766-11774.

[10] Chaves R, Ramírez J, Górriz JM, Illán IA. Functional Brain Image Classification Using Association Rules Defined Over Discriminant Regions. Pattern Recognit. Lett. 2012;33:1666-1672.

[11] Liu M, Zhang D, Shen D. Ensemble Sparse Classification of Alzheimer's Disease. NeuroImage. 2012;60:1106-1116.

[12] Westman E, Muehlboeck JS, Simmons A. Combining MRI and CSF Measures for Classification of Alzheimer's Disease and Prediction of Mild Cognitive Impairment Conversion. Neuroimage. 2012;62:229-238.

[13] Veeramuthu A, Meenakshi S, Manjusha PS. A New Approach for Alzheimer's Disease Diagnosis by Using Association Rule Over Pet Images. Int. J. Comput. Appl. 2014;91:9-14.

[14] Fulton LV, Dolezel D, Harrop J, Yan Y, Fulton CP. Classification of Alzheimer's Disease With and Without Imagery Using Gradient Boosted Machines and Resnet-50. Brain Sci. 2019;9:212.

[15] Lella E, Lombardi A, Amoroso N, Diacono D, Maggipinto T, et al. Machine Learning and DWI Brain Communicability Networks for Alzheimer's Disease Detection. Appl. Sci. 2020;10:934.

[16] Bae JB, Lee S, Jung W, Park S, Kim W, et al. Identification of Alzheimer's Disease Using a Convolutional Neural Network Model Based on T1-Weighted Magnetic Resonance Imaging. Sci. Rep. 2020;10:22252.

[17] Fathi S, Ahmadi A, Dehnad A, Almasi-Dooghaee M, Sadegh M. A Deep Learning-Based Ensemble Method for Early Diagnosis of Alzheimer's Disease Using MRI Images. Neuroinformatics. 2024;22:89-105.

[18] Liu C, Ouyang C, Cheng S, Shah A, Bai W, et al. G2d: From Global to Dense Radiography Representation Learning via Vision-Language Pre-training. 2023. ArXiv preprint: https://arxiv.org/pdf/2312.01522

[19] Jenber Belay A, Walle YM, Haile MB. Deep Ensemble Learning and Quantum Machine Learning Approach for Alzheimer's Disease Detection. Sci. Rep. 2024;14:14196.

[20] Sarma M, Chatterjee S. Etiology of Late-Onset Alzheimer's Disease, Biomarker Efficacy, and the Role of Machine Learning in Stage Diagnosis. Diagn. 2024;14:2640.

[21] Kienzler R. The Lightweight IBM Cloud Garage Method for Data Science. A Process Model to Map Individual Technology Components to the Reference Architecture. 2019.

[22] Zhai H, Sands T. Comparison of Deep Learning and Deterministic Algorithms for Control Modeling. Sens. 2022;22:6362.

[23] Arya M, Hanumat Sastry G. A Novel Deep Ensemble Learning Framework for Classifying Imbalanced Data Stream. In IOT with Smart Systems: Proceedings of ICTIS . 2022,251:607–617.

[24] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of artificial intelligence research (JAIR) . 2002;16:321-357.

[25] Fernández A, García S, Galar M, Prati RC, Krawczyk B, et al. Learning from Imbalanced Data Sets. Cham: Springer. 2018.

[26] He H, Ma Y. Imbalanced Learning: Foundations, Algorithms, and Applications. 2013.

[27] An N, Ding H, Yang J, Au R, Ang TF. Deep Ensemble Learning for Alzheimer's Disease Classification. J. Biomed. Inform. 2020;105:103411.

[28] Brownlee J. Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning. Machine Learning Mastery. 2020.

[29] Cárdenas AA, Baras JS. B-ROC Curves for the Assessment of Classifiers Over Imbalanced Data Sets. In Proceedings of the Twenty-First national conference on artificial intelligence. AAAI Press. 2006:21;1581-1584.

[30] Ferri C, Hernández-Orallo J, Flach PA. A Coherent Interpretation of Auc as a Measure of Aggregated Classification Performance. In Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011:657-664.

Appendix

Abbreviation	Definition
AD	Alzheimer's Disease
CN	Cognitive Normal
MCI	Mild Cognitive Impairment
EMCI	Early MCI
LMCI	Late MCI
ADNI	Alzheimer's Disease Neuroimaging Initiative
NACC	National Alzheimer's Coordinating Center
MRI	Magnetic Resonance Imaging
ROI	Region Of Interest
FDG	Fluorodeoxyglucose
PET	Positron Emission Tomography
DTI	Diffusion Tensor Imaging
CSF	Cerebrospinal fluid
APOE	Apolipoprotein E
MMSE	Mini Mental State Examination
CDR	Clinical Dementia Rating
ADAS	Alzheimer Disease Assessment Scale
RF	Random Forest
SVM	Support Vector Machines
ANN	Artificial Neural Network
SMOTE	Synthetic Minority Oversampling Technique
ML	Machine Learning
DL	Deep Learning
GBM	Gradient Boosted Machine
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic