Determination of Alzheimer's Disease Stages by Artificial Learning Algorithms

Nurgul Bulut[1,*], Tuna Cakar[2] , Ilker Arslan[3], Zeynep K. Akinci[4], Kevser S. Oner[5]

[1]*Department of Biostatistics and Medical Informatics, Faculty of Medicine, Istanbul Medeniyet University, 34000, Istanbul, Turkey*

[2]*Department of Computer Engineering, Faculty of Engineering, MEF University, 34396, Istanbul, Turkey*

[3]*Department of Mechanical Engineering, Faculty of Engineering, MEF University, 34396, Istanbul, Turkey*

[4]*Department of Neurology, Sultan Abdulhamid Han Research and Training Hospital, Saglik Bilimleri University, 34668, Istanbul, Turkey*

[5]*Department of Biostatistics, Faculty of Medicine, Eskisehir Osmangazi University, 26040, Eskisehir, Turkey*

Abstract: *Introduction:* This study aims to determine the stages of Alzheimer's disease (AD) using different machine learning algorithms, and compares the performance of these models.

Methods: Demographic, genetic, and neurocognitive inventory data from the National Alzheimer's Coordinating Center (NACC) database as well as brain volume/thickness data from magnetic resonance imaging (MRI) scans were used. Deep Neural Networks, Ordinal Logistic Regression, Random Forest, Gaussian Naive Bayes, XGBoost, and LightGBM models were used to identify four different ordinal stages of AD.

Results: Although the performance measures of the developed models were similar, the highest classification rate of AD stages was achieved by the Random Forest model (accuracy: 0.86; F1 score: 0.86; AUC: 0.95). The outputs of the model with the best performance were explained by the SHapley Addictive exPlanations (SHAP) method.

Conclusions: This indicates that non-invasive markers and machine learning models can be used effectively in early diagnosis and decision support systems to predict stages of AD.

Keywords: Alzheimer's Disease, Artificial Intelligence, National Alzheimer's Coordinating Center, Machine Learning, Explainable Artificial Intelligence, SHapley Addictive exPlanations, Artificial Learning Algorithm.

1. INTRODUCTION

Dementia, a condition that affects cognition and social skills, is one of the most rapidly spreading diseases in the world. The most common cause of dementia is Alzheimer's disease (AD), with a rate of approximately 60-80% [1]. Dementia and Alzheimer's disease, whose symptoms are worsened with age, are difficult to diagnose early and as of now, are not curable or reversible. According to the latest census report of the United Nations, people over the age of 65 constitute 10% of the world population, and that number is expected to rise to 16% by 2050 [2]. Parallel to this aging population trend, the prevalence of AD is increasing worldwide as well [2]. Once diagnosed with Alzheimer's disease, the average remaining lifespan of patients is 4-8 years, although this can extend up to 20 years depending on other factors [3].

The progression of Alzheimer's disease, which varies according to the age and health status of the person, is described as the atrophy and eventually death of brain cells. Brain changes begin in the years leading up to the first appearance of symptoms, including damage to neurons in other parts of the brain as the disease progresses [3, 4]. The disease, which begins to manifest itself with cognitive and behavioral disorders affecting the person's daily life activities, is categorized into four stages: questionable, mild, moderate, and severe [4]. Clinical findings, imaging methods, laboratory examinations, genetic tests, and neuropsychological inventories are used for diagnosis. Biomarkers also change throughout the stages of AD [5]. Neuropsychological assessments that can be performed in a short amount of time are used for the diagnosis of AD, particularly in the early stages. Genetic tests are also used as an aid to diagnosis, especially tests detecting the presence of ε2, ε3, and ε4 alleles of the Apolipoprotein E (*APOE*) genotype [5]. MRI provides detailed visualization of the size and shape of the brain and brain regions. Positron emission tomography (PET) uses small amounts of radioactive material called tracers to measure specific activity, such as glucose utilization, in different brain regions. Cerebrospinal fluid (CSF) can be measured both by

*Address correspondence to this author at the Department of Biostatistics and Medical Informatics, Faculty of Medicine, Istanbul Medeniyet University, 34000, Istanbul, Turkey; E-mail: nrgl.bulut@hotmail.com

taking fluid from brain samples and via MRI scans. CSF is used to measure amyloid β42 levels, amyloid β42/40 ratio (the main component of amyloid plaques in the brain), total tau and phosphorylated-tau protein levels (the main components of tau wrinkles in the brain), which are important biomarkers of Alzheimer's disease. The abnormal levels of amyloid deposits are also measured with amyloid PET or Tau PET scans [3]. Once the brain structure imaging is completed, the data obtained is employed in mathematical calculations such as those pertaining to volume, stroke volume, and insufficiency measurement. The development of diagnostic models continues to be a priority, with all imaging and volume values being utilized in this process [5].

Alzheimer's disease requires long-term care and causes social, economic, psychological, and physical problems not only to the patients themselves but also to their caregivers, their families, and to society. With the increase of the disease prevalence rate, many countries are creating prevention programs and plans, and using various methods for diagnosis [1, 3]. Although the number of people with AD in the aging world population is increasing, the underlying causes of the disease are not fully understood yet. Artificial intelligence-based studies on the subject and the opening of AD centers and databases to researchers contribute to the increase in research on the diagnosis and treatment of AD [6]. The main objective of this study is to determine the stages of AD using machine learning models trained/ based on the NACC dataset that could be proposed as a decision support system for physicians in the sense that it will help them to make better and faster decisions.

2. MATERIALS AND METHODS

2.1. Dataset

We used patient data from the NACC, which collects standardized clinical and neuropathological research data from Alzheimer's Disease Research Centers (ADRCs) in the United States, creates imaging databases, and develops Uniform Data Sets (UDS). Participants who applied to the ADRCs were patients and their relatives and were willing to participate in dementia research. Participants were recruited in accordance with the protocols of each center, and consent was obtained for data collection. Data were collected by clinicians or trained interviewers working in these centers [7, 8].

The data were selected from the UDS version of patients examined between June 9, 2005, and February 22, 2022, frozen in March 2022. The dataset of annual examination results filled in by specialists for the 45,100 ADRC patients followed at 37 different centers consists of electronic health records, digital biomarkers, digital neuropathology data, and genetic results in the UDS database and measurements from MRI overwritten images in the imaging database [7, 8].

2.2. Data Preprocessing

The dataset contains the results of a total of 166,082 observations from repeated examinations of 45,100 patients aged 18 to 110 years, who were each examined at least 1 and at most 17 times. In order to include patients with late (severe) stages of Alzheimer's disease, the most recent visit was selected from the repeated examinations, and a single observation result was retained from each patient. The dependent variable was the global Clinical Dementia Rating (CDR) score, which grades the Alzheimer's stage in patients over 45 years of age. CDR scale ratings are graded as 0 = healthy, 0.5 = questionable, 1 = mild, 2 = moderate, and 3 = severe [9]. The number of patients with CDR scale ratings 2 and 3 was lower than the other groups. In our study, "moderate" and "severe" stages were combined to eliminate the imbalance of distribution between levels; similarly, to avoid confusion with decimal numbers, the value "0.5" was changed to "1." The final CDR scale ratings were therefore graded as 0 = healthy, 1 =questionable, 2 = mild, 3 = moderate and severe.

Of all the variables in the dataset, we examined the significance and association stages with the CDR scale ratings, and those found to be significant were selected as independent variables. All patients took a test, either the Mini-Mental State Examination (MMSE), used until 2015, or the Montreal Cognitive Assessment (MoCA), rolled out in 2014. MoCA scores were converted to MMSE scores to compensate for the loss of information [10, 11]. The relevant score of the patient was calculated by summing the variables of the 10-section Functional Activities Questionnaire (FAQ) inventory, which examines the cognitive impairment status of the patients. The results of 12 different conditions considered by the Neuropsychiatric Inventory (NPI), which measures the severity of neuropsychiatric symptoms, were selected from the dataset, providing us with a total NPI score. Geriatric Depression Scale (GDS) scores indicating the result of depression screening were included in the dataset. To confirm the

diagnosis of the disease, *APOE* genotype analysis results were selected from the gene dataset and added as an independent variable.

Although patients did not have MRI scans every time they were examined, 11,273 MRI results were available for a total of 7,328 patients, with repeat MRI scans for some patients. From the UDS and imaging databases, we selected patients with the highest CDR and at different AD stages, with a maximum of 180 days between examination dates [12, 13]. The dataset contains a total of 155 measured values calculated from different regions of the brain. In order to correct the individual variability of brain structure measurements, normalization was achieved by dividing all measurement values by the total brain volume [14]. The significance of the normalized volumetric and thickness calculations was examined according to the CDR scale rating. The highly correlated measurements were eliminated, and the deviant values of the remaining 17 measurements were removed from the dataset. Missing data were filtered according to the independent variables identified. When patients with physical, cognitive, and behavioral health problems were excluded, a dataset of 1,543 patients (healthy = 967; questionable= 409; mild = 127; moderate-severe = 40) was obtained.

2.3. Data Analysis

In this study, we evaluated the performances of six models: Ordinal Logistic Regression (LR), Random Forest (RF), Gaussian Naive Bayes (NB), XGBoost, LightGBM, and Deep Neural Network (DNN). The dataset was converted to values between 0 and 1 by the min-max normalization method. The min-max normalization method, which was used to remove bias and equalize distances between variables, as the data were not normally distributed. For the DNN model, 70% of the dataset was allocated for training, 15% for validation, and 15% for testing, while for the other five models, 75% was allocated for training and 25% for testing. To address the high dimensionality of the MRI-derived variables, we employed a two-step feature reduction strategy. First, we used Locally Linear Embedding (LLE) to project the original 155 volumetric and thickness measures onto a lower-dimensional manifold, capturing the most relevant non-linear structures in the data. After experimenting with different dimensionality settings (ranging from 5 to 15), we chose 7 dimensions based on the optimal trade-off between retaining variance and achieving high classification accuracy.

In the second step, we utilized Recursive Feature Elimination (RFE) across the full predictor set—including both MRI-derived features (transformed via LLE) and non-imaging variables such as the FAQ, MMSE, and APOE genotype. RFE systematically ranks features by their importance to the predictive model, iteratively pruning the least influential features. This approach ensured that only a parsimonious subset of high-impact features was retained for the final training and optimization. Empirically, we observed that increasing the number of features beyond approximately 25 did not yield further improvement in the model's test accuracy, indicating that our feature reduction approach effectively mitigated overfitting and redundancy in the predictors.

The results of LLE algorithm, which is one of the methods of reducing the multivariate measurements obtained from MRI scanners to a lower dimension by capturing high-dimensional non-linear features, were optimized and included in the models [15]. In addition to LLE, principal component analysis (PCA) and ISOMAP were also used for dimensionality reduction, but LLE was used because the data structure is both non-linear and performs better. The random grid search method was used in all model development processes to prevent overlearning and to determine which hyperparameters would provide the best model performance.

Synthetic Minority Over-sampling Technique (SMOTE), an advanced resampling technique used to solve class imbalance problems, was applied. Sensitivity, specificity, precision, F1 score, accuracy, and Area Under the Curve (AUC) were calculated to compare the performance of the models. Accuracy gives a quick overview, but the F1 score helps balance precision and recall. For unbalanced data, sensitivity and specificity help understand how the model handles positive and negative classes. AUC shows how well the model separates classes, useful for choosing the right decision threshold. Contingency matrix and receiver operating characteristic curve (ROC) graph outputs were obtained. The features of the model with the best performance were explained with the SHAP analysis. Finally, the effect of the number of features selected by RFE on model test accuracy was plotted.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives}$$

Data analysis was performed using the Python programming language (version 3.10) and the cloud-based Google Colaboratory (colab) environment, as well as the Python libraries NumPy 1.23.0, Matplotlib 3.6.0, Scikit-Learn 1.1.3, TensorFlow 2.10, Statsmodel 0.13.5, and Pandas 1.5.0.

3. RESULTS

The model evaluation indicated that in the training dataset, the highest F1 score was achieved by XGBoost with SMOTE (0.89), followed by other models like DNN (0.86), LightGBM (0.86), and RF (0.86) when SMOTE was applied. Ordinal Logistic Regression (LR) and Gaussian Naïve Bayes (NB) performed better without SMOTE (0.84 and 0.83, respectively). The highest accuracy was also observed in XGBoost with SMOTE (0.89). DNN and LightGBM had improved accuracy with SMOTE, whereas RF maintained consistent accuracy (0.86) regardless of SMOTE. In the test dataset, Random Forest (RF) without SMOTE had the highest F1 score (0.86), outperforming XGBoost (0.85) and DNN (0.84). LR and NB had better F1 scores without SMOTE (0.85 and 0.85, respectively), while LightGBM performed the worst (0.83). Accuracy results followed a similar trend, with RF maintaining a stable accuracy of 0.86 and LightGBM showing the lowest accuracy (0.82–0.83). Overall, Random Forest emerged as the most effective model in the test dataset, excelling in both F1 score and accuracy. LightGBM was the least effective model based on these metrics. The best hyperparameters for RF: criterion='entropy', max_depth=5, max_features=7, min_samples_leaf=10, min_samples_split=13, n_ estimators=500.

The performance metrics of the Random Forest model, which performed the best in the test dataset, are presented in Table **2** for 4 different classes. When these results are evaluated, it is worth pointing out that the performance is higher at the extremes (classes 0 and 3), whereas for the classes in between, the performance for different metrics tends to be lower.

Figure **1** presents the confusion matrices of CDR scale ratings and compares the performance of the standard and SMOTE-implemented versions of the Random Forest algorithm. Accordingly, out of 242 healthy individuals, 231 were correctly predicted according to the standard RF model, while 213 were correctly predicted when SMOTE was applied. In the first matrix (RF), a relatively high correct classification rate (231 correct, 11 incorrect) is observed for class 0.

Table 1: F1 Score and Accuracy Metrics of the Models

	Training				Test			
	NonSMOTE		SMOTE		NonSMOTE		SMOTE	
	F1 score	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score	Accuracy
DNN	0.82	0.83	0.86	0.85	0.83	0.84	0.84	0.85
LR	0.84	0.83	0.81	0.81	0.85	0.84	0.79	0.80
RF	0.86	0.86	0.86	0.86	**0.86**	**0.86**	0.83	0.83
Gaussian NB	0.83	0.82	0.80	0.80	0.85	0.85	0.84	0.84
XGBoost	0.88	0.87	0.89	0.89	0.84	0.84	0.85	0.85
LightGBM	0.84	0.83	0.86	0.85	0.83	0.82	0.83	0.83

Table 2: Performance of the RF Model at CDR Scale Rating

	CDR	NonSMOTE					
		Sensitivity	Specificity	Precision	F1 score	Accuracy	AUC
RF	**0**	0.88	0.91	0.95	0.92	0.89	0.94
	1	0.80	0.88	0.66	0.72	0.87	0.88
	2	0.79	0.98	0.81	0.80	0.97	0.98
	3	1.00	0.99	0.70	0.82	0.99	1.00

AUC = Area under the curve.

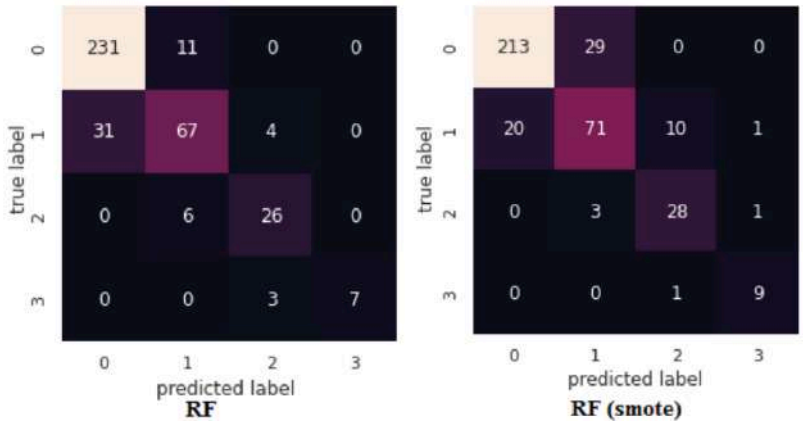

Figure 1: RF model CDR scale rating confusion matrices.

However, the correct prediction rates for classes 2 and 3, which are undersampled due to imbalance, are significantly lower. In particular, class 3 is limited to only seven correct predictions. In the second matrix (RF-SMOTE), with the application of SMOTE, the class imbalance was eliminated, and the correct prediction rates increased in classes with low number of samples (classes 2 and 3). The number of correct classifications for class 3 increased to 9. However, the number of incorrect predictions (29 incorrect) increased for class 0. This suggests that while SMOTE improves performance in the minority classes, it may cause a slight performance degradation in the majority classes. Overall, the application of SMOTE made the model performance more balanced by reducing the effects of class imbalance.

In addition, the graphs showing the loss and accuracy values at each epoch for the training and validation datasets in the DNN model are presented in Figure **2**. It was found that the loss of the training and validation datasets were almost equal and less than 1 for 100 epochs. In the model, the accuracy values of the training and validation datasets were between 0.75 and 0.85 for 100 epochs. In the model with SMOTE, the training and validation datasets were very close to each other after 40 epochs, and both were at 0.80 and above towards the end. The training and validation loss started at a very high stage at the beginning and decreased rapidly. This shows that the model starts to learn the patterns in the dataset effectively at an early stage. From around epoch 20 onwards, both training and validation losses slow down to a steady decline. The fact that the training and validation curves are close to each other indicates that the model has a good generalization capacity and does not suffer from overfitting. Moreover, the low stage of validation loss indicates that the model can also be successfully applied on test data. These results indicate that the training process of the model is managed efficiently, and the hyperparameter choices are appropriate.

Among the 155 measurements obtained from MRI scanners, 17 measurements were optimized with LLE algorithms, reduced to 7 dimensions, and included in the model. For the RF model, which performed best of all the models tested, SHAP analysis revealed that the variables with the highest contribution to the model were, in decreasing order, FAQ, MMSE, NPI, and INDEPEND, as shown in Figure **3**. The SHAP results revealed that LLE_1, that included total cerebrospinal fluid volume (cm^3) and left cuneus mean cortical

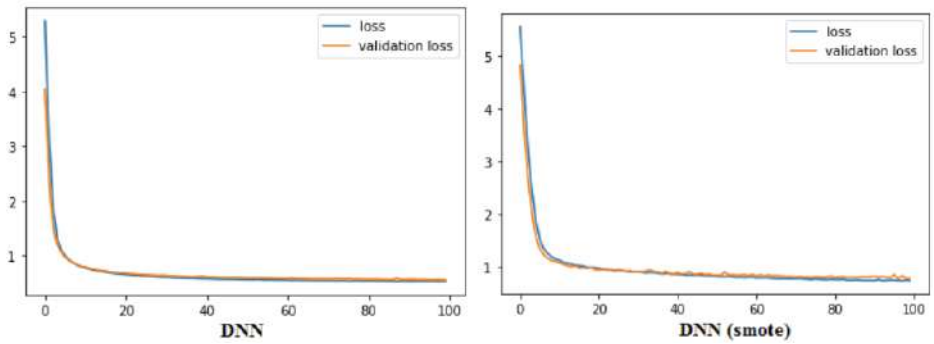

Figure 2: Training/Validation loss of DNN model.

thickness (mm), and LLE_5 that included total cerebrum volume (cm^3) and left entorhinal mean cortical thickness (mm) measurements.

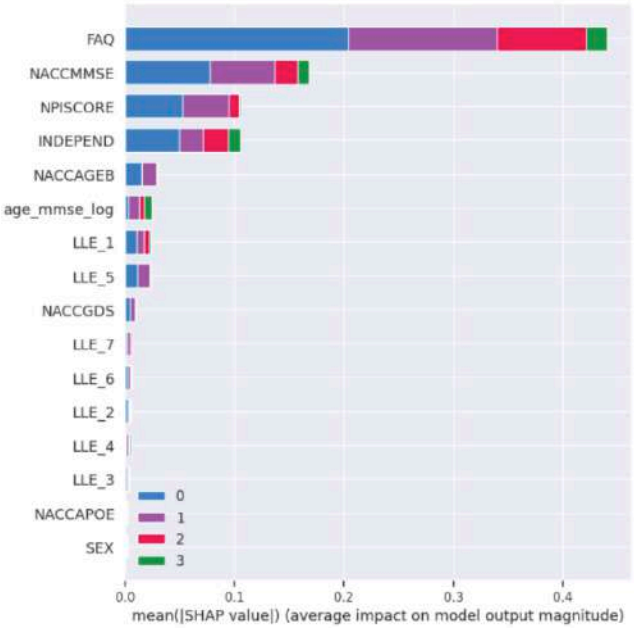

Figure 3: SHAP values for the Random Forest model.

The recursive feature elimination method was applied on the total feature set of 155 measurements and reduced to 22 MRI measurements. Then the RF model was developed after the LLE method was applied on these 22 MRI measures and the accuracy of the RF model was calculated as 0.89 for training and 0.82 for testing. The effect of the number of features selected with the RFE method on the test accuracy of the model is shown in Figure **4**. The graph reveals that initially, with a small number of features, the model

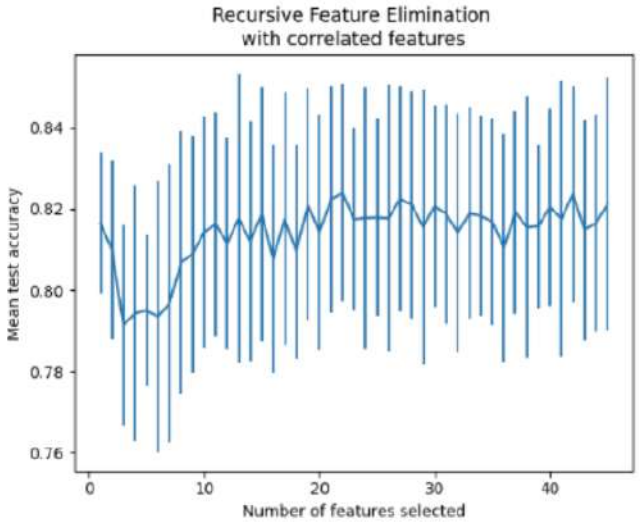

Figure 4: The RF model performance with respect to the recursive feature elimination method.

accuracy is low (0.78). As the number of features increases, there is a clear increase in accuracy, reaching a maximum level of accuracy between approximately 10-25 features (0.84). This is a critical region for determining the optimal number of features. Beyond 30 features, the accuracy stabilizes and the variance increases, indicating that adding more features to the model does not contribute to better performance and instead creates excessive complexity. While the correlation between features may explain the initial low accuracy values, it appears that RFE effectively eliminates redundant features and optimizes the model.

4. DISCUSSION AND CONCLUSION

The main topic of this study is the use of machine learning models to determine different stages of AD, the most common cause of dementia and an irreversible neurodegenerative disease. For diagnosis, clinical findings, imaging methods, laboratory examinations, genetic tests, and neuropsychological inventories help detect the disease in machine learning models. For this purpose, a dataset of 45,100 patients (healthy = 15,837; questionable= 13,154; mild = 6,774; moderate-severe = 9,335) from the NACC database was used, and exclusionary factors were eliminated. The results of the clinical inventories, neurocognitive assessments, and the measurements of MRI images were evaluated in the added dataset, composing only about 3.42% of the entire database. It has been proven that women are twice as likely to have AD compared to men due to a longer life expectancy, higher rates of depression, lower education levels, lower stages of estrogen (the hormone that protects the mental acuity of the brain after menopause), and a stronger *APOE* ε4 genotype [16]. In our study, the difference between genders was eliminated by proportioning the MRI measurements of the patients to their whole brain volumes. Other factors known to affect Alzheimer's disease are marital status, family history, smoking and alcohol use, cardiovascular risk factors, diabetes, hypertension, cholesterol, obesity, and head trauma. We could not include these variables in our models for two reasons: first, although some of them are significant for the diagnosis of AD, they do not contribute to its classification; second, when all of these variables were included in the model, the number of participants meeting the required criteria dropped to zero due to missing data, leaving no usable sample for analysis.

In this framework, LR, RF, Gaussian NB, XGBoost, LightGBM, and DNN models were developed and

optimized by fine-tuning for multiple comparisons of four different ordinal stages of AD. The data used for the optimization were demographic, genetic, and neurocognitive inventory results of patients from the NACC database and brain volume/thickness measurements calculated from MRI scanners. In the models, divided between training/validation and test datasets, LLE (a dimensionality-reduction method for MRI measurements) and SMOTE techniques were used to overcome the problem of unbalanced distribution of the number of samples in each class. Although the models used are popular algorithms with good predictive power and processing speed, better results were obtained when the parameters were adjusted according to the data structure.

The findings of our study revealed that the accuracy rate, F1 score, AUC value and sensitivity, specificity, and precision performance measures of each class of the developed models were similar and that it was the RF model without SMOTE. In addition, the contribution of the variables in the RF model in determining AD stages was explained by SHAP analysis. Accordingly, it was found that the most successful classifier of AD stages among the variables included in the model was the FAQ inventory. In our model, reduced from 155 to 17 measurements using LLE algorithms, the most successful MR measurements in AD stages classification were found to be total cerebrospinal fluid volume, left cuneus mean cortical thickness, total cerebrum volume, and left entorhinal mean cortical thickness. In addition, a higher performance model was established by applying the dimension reduction method to the multidimensional dataset. In conclusion, the machine learning-based models employed in this study can be used to identify patients suspected of having AD in the early diagnosis process with cost-effective and widely-used, non-invasive markers to predict at what stage of AD they may be. These models are generalizable, as the findings obtained show similar stages of performance for training and test datasets.

Several other studies in the academic literature used machine learning techniques to diagnose or classify AD using both clinical inventory results and imaging methods. Yang *et al*. normalized the volumetric measurements of MRI and CSF images of 200 healthy, 400 questionable and 200 Alzheimer's disease patients from the Open Access Series of Imaging Studies (OASIS) and Alzheimer's Disease Neuroimaging Initiative (ADNI) databases, then applied dimensionality reduction with the independent component analysis (ICA) method and compared the performance of the Support Vector Machine (SVM) algorithm to binarily discriminate the questionable and patient classes from the healthy class. Accordingly, the SVM algorithm was able to discriminate between the healthy from the questionable class with a maximum accuracy of 0.81 and from the patient class with 0.89 [17].

In another study, Zhang *et al*. proposed to combine MRI, PET, and CSF biomarkers from the ADNI database, and introduced a multimodal data fusion and classification method using a multicore SVM classifier to distinguish the patient or questionable class from the healthy class, integrating the three methods for the classification task [18]. A simple feature selection based on a t-test was used to select the most discriminative features; using the volumetric features extracted from the images, the best binary classification rate they obtained was 0.93 (sensitivity: 0.93 and specificity: 0.93) for the SVM model with 10-fold cross-validation. In another study, Wolz *et al*. compared the performance of two-class models built with Linear Discriminant Analysis (LDA) and SVM algorithm using age, gender, education level, MMSE, GDS, and MRI measurements of 66 healthy and 48 AD patients from the same database. They found the LDA model performed best in the classification of the healthy group versus the patient group (accuracy; 0.89; sensitivity: 0.93; specificity: 0.85) [19].

Liu *et al*. designed a study to classify Alzheimer's patients with DNN architecture using MRI and PET data from imaging methods. After processing the images in the ADNI database, they divided the original dataset into two parts: 90% training set and 10% test set. They calculated the performance metrics of the trained network for both binary and multiclass classification by selecting the number of neurons between 30 and 200 for two hidden layers using the softmax activation function in each layer and the relative weights of each nucleus by grid search with a learning rate of 0.1 at each step. They achieved an accuracy of 0.82 for the questionable class and 0.91 for the patient class compared to the healthy class, but only 0.54 for multiclass classification [20]. In another study, Ritter *et al*. constructed a dataset including demographics, neuropsychological tests (MMSE, GDS, NPI, ADAS, FAQ), medical history, baseline medical symptoms, genes, amyloid plaques, neurological and physical examinations, MRI, FDG-PET, CSF from the ADNI database of patients who did or did not convert to AD within 3 years. Missing data was completed using mean value and expectation maximization algorithms. With 10-fold cross-validation SVM, Decision Tree, and

RF models, they achieved 0.73 accuracy (0.40 sensitivity and 0.91 specificity) and overall, the highest score obtained by the SVM algorithm for binary classification into AD [21].

Khagi *et al*. used MRI images of 18 healthy, 16 questionable, 12 mild, and 4 moderate Alzheimer's patients from the OASIS database to classify the images using a simple machine learning algorithm with deep layers feature extraction using Deep Neural Network architecture. After extracting the features of the images with the CNN layer, they used Mutinffs, ReliefF, Laplacian, and UDFS algorithms for feature selection. With K-Nearest Neighbor and SVM machine learning techniques, they obtained an accuracy of 0.99 in the multiclass model with five-fold cross-validation in the dataset, which was divided into 70% training and 30% test set [22]. Liu *et al*. evaluated the MRI data of 119 healthy, 233 questionable, and 97 AD patients from the ADNI database. They combined the DesneNet model for learning image features based on segmented hippocampal regions with a multi-task CNN model for learning hippocampal segmentation and classifying disease status. With the features learned from these models, they correctly predicted questionable against healthy people with a rate of 0.76, and patients with a rate of 0.89 [23]. In a subsequent study, Liu *et al*. extracted an MRI dataset of 492 participants aged 18-96 years from the OASIS database and developed a CNN model trained with Alexnet and GoogLeNet networks, and predicted healthy, questionable, and patient classes with an accuracy of 0.78, sensitivity of 0.83, and specificity of 0.75 with multiple comparisons [24].

Using the NACC database, González *et al*. compared the AUC values of the models they developed based on RF algorithms for healthy, mild cognitive impairment, and dementia groups for 7,054 participants with age, gender, education level, ethnic group, language, FAQ, and MoCA variables [25]. They optimized the number of trees and the number of attributes to split each node by dividing 20% (1,410) of the primary dataset into a test set and 80% (5,644) into a training set. They found the AUC of the highest-performing model, including racial disparities corrected for demographic variables, to be 0.88 [25]. They predicted the highest diagnostic accuracy for the healthy (AUC = 0.91) and mild cognitive impairment (AUC = 0.84) groups, which they compared with other classes. The model they found to perform best was more successful in classifying dementia patients than those in the healthy group and more successful in classifying the mild cognitive impairment group than the findings in González *et al*. [25]. They emphasized the difficulty of separating the group with mild cognitive impairment from the healthy group, which resulted in misclassifying 217 people with mild cognitive impairment in their model and predicting that they were healthy [25].

In order to illustrate the robustness of our final models, Table **3** provides a comparative view of our best-performing classifier—Random Forest (RF)—against empirical results from selected recent studies in the literature that used MRI-derived and clinical data to classify Alzheimer's disease stages. The table displays accuracy, F1 score, and AUC values to facilitate direct comparison. As seen in Table **3**, our final RF model yields favorable results—an accuracy of 0.86 and an F1 score of 0.86—while maintaining an AUC of 0.95. These findings align closely with or surpass other approaches in the literature, especially considering the complexity of the four-class classification task in our study. Our results underscore both the efficacy of the proposed feature reduction technique and the enhanced diagnostic precision contributed by integrating clinical, genetic, and neuroimaging information.

Manca *et al*. investigated the neurocognitive effects associated with neuropsychological symptoms in

Table 3: Comparison of our Final Random Forest Model with Selected Prior Work on MRI-Based and/or Clinical Classification of Alzheimer's Disease

Study	Data Source	Model	Stages	Accuracy	F1 Score	AUC
Yang *et al*. (2011)	OASIS & ADNI	SVM	Binary (Healthy vs. AD / MCI)	0.81–0.89	–	–
González *et al*. (2021)	NACC	Random Forest	Binary (Healthy vs. MCI/Dementia)	0.82–0.91	–	0.84–0.88
Khagi *et al*. (2019)	OASIS	CNN + KNN/SVM	Multiclass (Healthy, MCI, AD)	0.99	–	–
Our Study	NACC	Random Forest (nonSMOTE)	Multiclass (Healthy, 3 AD stages)	0.86	0.86	0.95

MCI: Mild Cognitive Impairment; AD: Alzheimer's Disease; SVM: Support Vector Machine; CNN: Convolutional Neural Network.

sexual minority groups using the NACC database. They developed multivariate general linear models for cognitively healthy and impaired groups based on variables related to brain structures and cognitive performance in participants over 55 years of age [11]. Using MRI brain volume and thickness, total intracranial volume, total cerebrospinal fluid volume, total white matter volume, gray matter regions, volume/thickness measurements of frontal, parietal, occipital, and temporal cortex, MMSE and NPI inventories measuring cognitive performances, and demographic characteristics such as gender, education level, and *APOE* genotype, they showed that individuals with cognitive impairment in the same-sex relationship group had significantly smaller parahippocampal volumes than healthy individuals. Another significant discovery was that individuals in same-sex relationships had better episodic memory. In their study, they suggested that the observed differences in cognitive outcomes between relationship groups may be influenced by unexplored protective factors against cognitive decline rather than the type of relationship itself. They highlighted social support as a potential protective factor, warranting further investigation in future studies [11].

Data from the NACC dataset used in our study was classified as healthy, questionable, mild, and moderate-severe stages of AD using six different machine learning models. For the first time in the literature, this dataset was applied to a multiclass classification method using a variety of machine learning algorithms with both clinical and MRI measurements. The models were compared with the performance metrics of the test sets. Using our proposed models based on the combination of cognitive and functional biomarkers and MRI measurements, we obtained successful results that will contribute to the literature on multiclass classification. The success rates of our models ranged from 0.83 to 0.86, while the AUC results ranged from 0.92 to 0.96. The highest F1 score and accuracy rate were observed in the RF model, both with a rate of 0.86. Random Forest often achieves high accuracy in our test data because it averages multiple decision trees, each trained on randomly sampled subsets of both data and features. This “bagging” and randomization reduce overfitting by ensuring the individual trees are less correlated. Additionally, Random Forest naturally models non-linear relationships without the need for extensive feature engineering, making it robust to noise and outliers. In our study, careful hyperparameter tuning—particularly regarding the number and depth of trees—further optimized performance. Finally, synergizing Random Forest with feature reduction (via Locally Linear Embedding and recursive feature elimination) minimized redundant variables, improving both the model’s interpretability and predictive power. In the same model, the accuracy rates for each stage of AD were 0.89, 0.87, 0.97, and 0.99, respectively. In the confusion matrices results of the models, it was observed that the number of incorrect predictions in all classes was both lower and closer to the limits. The contributions of the variables in our model to the model were also made more explainable by the SHAP analysis. Thus, it was shown more clearly which variable contributed to determining the stage of AD.

However, as mentioned in the results section, the cases where the models give erroneous outputs are almost always in neighboring classes. Especially in diseases such as Alzheimer's disease, which is composed of many factors, high transitivity between classes is an expectable situation. On the other hand, the imbalance between classes seen in the dataset is one of the main limitations encountered in the model development process. In the future, in addition to the data augmentation methods used in this study, deep learning algorithms will be used for MRI and PET images. In addition, since the diagnoses become clearer during follow-up, longitudinal data of the same repeated patients will be used to predict changes in Alzheimer's stages using different machine learning techniques. When the findings and performance values obtained in this study are compared with other studies on the same dataset, two contributions are worth pointing out. First of all, the AUC values obtained in this study are higher compared to the outputs of the models developed by González *et al*. [24]. The main reason for this may be that the filters used in our study were used more effectively, and outliers were eliminated. Second, when the deep learning architecture performances used by Olaimat *et al*. are compared with the results of our study, it is seen that the model performances we developed are higher [26].

In conclusion, Alzheimer's disease is a significant public health concern worldwide, requiring long-term care and increasing the responsibilities of caregivers as it gradually worsens. Many methods are used for diagnosis, and machine learning models are also used to detect and classify the disease. The main objective of this study is to develop successful models for disease detection with machine learning algorithms by using different feature transformation and engineering

methods in data science with volumetric, demographic, and other data of healthy, questionable, and Alzheimer's patients. In this context, our study contributes to the academic literature in terms of the performance of machine learning models developed outside of attribute engineering processes.

DATA AVAILABILITY

The datasets used and analyzed during the current study are available from the NACC.

FUNDING SOURCES

The authors did not receive support from any organization for the submitted work.

COMPETING INTERESTS

The authors declare that they have no competing interests.

ACKNOWLEDGEMENTS

The NACC database is funded by NIA/NIH Grant U24 AG072122. NACC data are contributed by the NIA-funded ADRCs: P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez, MD), P30 AG062421 (PI Bradley Hyman, MD, PhD), P30 AG066509 (PI Thomas Grabowski, MD), P30 AG066514 (PI Mary Sano, PhD), P30 AG066530 (PI Helena Chui, MD), P30 AG066507 (PI Marilyn Albert, PhD), P30 AG066444 (PI David Holtzman, MD), P30 AG066518 (PI Lisa Silbert, MD, MCR), P30 AG066512 (PI Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI Andrew Saykin, PsyD), P30 AG072975 (PI Julie A. Schneider, MD, MS), P30 AG072978 (PI Ann McKee, MD), P30 AG072977 (PI Robert Vassar, PhD), P30 AG066519 (PI Frank LaFerla, PhD), P30 AG062677 (PI Ronald Petersen, MD, PhD), P30 AG079280 (PI Jessica Langbaum, PhD), P30 AG062422 (PI Gil Rabinovici, MD), P30 AG066511 (PI Allan Levey, MD, PhD), P30 AG072946 (PI Linda Van Eldik, PhD), P30 AG062715 (PI Sanjay Asthana, MD, FRCP), P30 AG072973 (PI Russell Swerdlow, MD), P30 AG066506 (PI Glenn Smith, PhD, ABPP), P30 AG066508 (PI Stephen Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, PhD), P30 AG072931 (PI Henry Paulson, MD, PhD), P30 AG066546 (PI Sudha Seshadri, MD), P30 AG086401 (PI Erik Roberson, MD, PhD), P30 AG086404 (PI Gary Rosenberg, MD), P20 AG068082 (PI Angela Jefferson, PhD), P30 AG072958 (PI Heather Whitson, MD), P30 AG072959 (PI James Leverenz, MD).

The authors are grateful to Dr. Konogan Beaufay for his extensive language check and editing.

ETHICAL APPROVAL

Not applicable.

CONSENT FOR PUBLICATION

Not applicable.

REFERENCES

[1] Alıcılar HE, Çalışkan D. Alzheimer's disease and prevention strategies. J Contin Med Educ 2021; 30(2): 107-15. https://doi.org/10.17942/sted.888837

[2] United Nations. World Population Prospects 2024: Summary of Results [Internet]. New York: United Nations; 2024 [cited 2024]. Available from: https://www.un.org/development/desa/pd/content/World-Population-Prospects-2024.

[3] Alzheimer's Association 2024 Alzheimer's disease facts and figures. Alzheimers Dement 2024; 20(5): 3708-21. https://doi.org/10.1002/alz.13809

[4] Mavioğlu H. Alzheimer's disease. In: Behavioral Neurology. 2nd ed. Istanbul: Nobel Medical Bookstores 2020; pp. 145-72.

[5] Jack CR, Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, *et al*. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol 2010; 9(1): 119-28. https://doi.org/10.1016/S1474-4422(09)70299-6

[6] Hoşgör H, Güngördü H. A qualitative research on the uses of artificial intelligence in health. Eur J Sci Technol 2022; (35): 395-407. https://doi.org/10.31590/ejosat.1052614

[7] Besser LM, Kukull WA, Teylan MA, Bigio EH, Cairns NJ, Kofler JK, *et al*. The revised National Alzheimer's Coordinating Center's Neuropathology Form-available data and new analyses. J Neuropathol Exp Neurol 2018; 77(8): 717-26. https://doi.org/10.1093/jnen/nly049

[8] Beekly DL, Ramos EM, Lee WW, Deitrich WD, Jacka ME, Wu J, *et al*. The National Alzheimer's Coordinating Center (NACC) database: the uniform data set. Alzheimers Dis Assoc Disord 2007; 21(3): 249-58. https://doi.org/10.1097/WAD.0b013e318142774e

[9] Gürvit İH, Baran B. Scales in dementias and cognitive disorders. Arch Neuropsychiatry 2007; 44(2): 58-65.

[10] Roalf DR, Moberg PJ, Xie SX, Wolk DA, Moelter ST, Arnold SE. Comparative accuracies of two common screening instruments for classification of Alzheimer's disease, mild cognitive impairment, and healthy aging. Alzheimers Dement 2013; 9(5): 529-37. https://doi.org/10.1016/j.jalz.2012.10.001

[11] Manca R, Correro AN, Gauthreaux K, Flatt JD. Divergent patterns of cognitive deficits and structural brain alterations between older adults in mixed-sex and same-sex relationships. Front Hum Neurosci 2022; 16: 620. https://doi.org/10.3389/fnhum.2022.909868

[12] Kovacevic S, Rafii MS, Brewer JB. High-throughput, fully-automated volumetry for prediction of MMSE and CDR

decline in mild cognitive impairment. Alzheimers Dis Assoc Disord 2009; 23(2): 139.
https://doi.org/10.1097/WAD.0b013e318192e745

[13] Tangaro S, Fanizzi A, Amoroso N, Bellotti R, Alzheimer's Disease Neuroimaging Initiative. A fuzzy-based system reveals Alzheimer's disease onset in subjects with mild cognitive impairment. Phys Med 2017; 38: 36-44.
https://doi.org/10.1016/j.ejmp.2017.04.027

[14] Mungas D, Reed BR, Jagust WJ, DeCarli C, Mack WJ, Kramer JH, *et al*. Volumetric MRI predicts the rate of cognitive decline related to AD and cerebrovascular disease. Neurology 2002; 59(6): 867-73.
https://doi.org/10.1212/wnl.59.6.867

[15] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science 2000; 290(5500): 2323-6.
https://doi.org/10.1126/science.290.5500.232

[16] Yang H, Oh CK, Amal H, Wishnok JS, Lewis S, Schahrer E, *et al*. Mechanistic insight into female predominance in Alzheimer's disease based on aberrant protein S-nitrosylation of C3. Sci Adv 2022; 8(50): eade0764.
https://doi.org/10.1126/sciadv.ade0764

[17] Yang W, Lui RL, Gao JH, Chan TF, Yau ST, Sperling RA, *et al*. Independent component analysis-based classification of Alzheimer's disease MRI data. J Alzheimers Dis 2011; 24(4): 775-83.
https://doi.org/10.3233/JAD-2011-101371

[18] Zhang D, Wang Y, Zhou L, Yuan H, Shen D, Alzheimer's Disease Neuroimaging Initiative. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 2011; 55(3): 856-67.
https://doi.org/10.1016/j.neuroimage.2011.01.008

[19] Wolz R, Julkunen V, Koikkalainen J, Niskanen E, Zhang DP, Rueckert D, *et al*. Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. PLoS One 2011; 6(10): e25446.
https://doi.org/10.1371/journal.pone.0025446

[20] Liu S, Liu S, Cai W, Che H, Pujol S, Kikinis R, *et al*. Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Trans Biomed Eng 2014; 62(4): 1132-40.
https://doi.org/10.1109/TBME.2014.2372011

[21] Ritter K, Schumacher J, Weygandt M, Buchert R, Allefeld C, Haynes JD, *et al*. Multimodal prediction of conversion to Alzheimer's disease based on incomplete biomarkers. Alzheimers Dement Diagn Assess Dis Monit 2015; 1(2): 206-15.
https://doi.org/10.1016/j.dadm.2015.01.006

[22] Khagi B, Kwon GR, Lama R. Comparative analysis of Alzheimer's disease classification by CDR level using CNN, feature selection, and machine-learning techniques. Int J Imaging Syst Technol 2019; 29(3): 297-310.
https://doi.org/10.1002/ima.22316

[23] Liu M, Li F, Yan H, Wang K, Ma Y, Shen L, *et al*. A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. Neuroimage 2020; 208: 116459.
https://doi.org/10.1016/j.neuroimage.2019.116459

[24] Liu J, Li M, Luo Y, Yang S, Li W, Bi Y. Alzheimer's disease detection using depthwise separable convolutional neural networks. Comput Methods Programs Biomed 2021; 203: 106032.
https://doi.org/10.1016/j.cmpb.2021.106032

[25] González DA, Gonzales MM, Jennette KJ, Soble JR, Fongang B. Cognitive screening with functional assessment improves diagnostic accuracy and attenuates bias. Alzheimers Dement Diagn Assess Dis Monit 2021; 13(1): e12250.
https://doi.org/10.1002/dad2.12250

[26] Al Olaimat M, Martinez J, Saeed F, Bozdag S. PPAD: A deep learning architecture to predict progression of Alzheimer's disease. bioRxiv 2023.
https://doi.org/10.1101/2023.01.28.526045