VER
Vascular & Endovascular Review

Towards Interpretable Alzheimer's Disease Diagnosis: A Multi-Model Machine Learning Framework with Consensus Feature Selection and SHAP-Driven Explainability

Fatima Tauseef[1], Ahmad Jamal[2], Fahad Naseer[3], Md Shafayet Hossain[4], and Dr. Mahzabin Mahmud[5]

[1]School of Business, Washington University of Science and Technology, 2900, Eisenhower Ave, Alexandria, VA 22314, USA
[2]King Graduate School, Monroe University, 434 Main St, New Rochelle, NY 10801,USA
[3]School of Information Technology, Washington University of Science and Technology, 2900 Eisenhower Ave, Alexandria, VA 22314, USA
[4]MSIT, Washington University of Science and Technology (WUST), 2900, Eisenhower Ave, Alexandria, VA 22314, USA
[5]Consultant Sonologist, Lab4 General Hospital, Rasel Bhaban, Ati Bazar-Kalatiya Rd, Dhaka-1312, Bangladesh
ajamal7503@monroeu.edu

ABSTRACT

Alzheimer's disease (AD) is a progressive neurodegenerative disorder representing the leading cause of dementia worldwide. Timely and accurate diagnosis remains a persistent clinical challenge owing to symptom overlap, diagnostic subjectivity, and the inaccessibility of neuro imaging-dependent workflows in resource-limited settings. This study proposes a comprehensive and clinically interpretable machine learning framework for automated AD classification using structured tabular clinical data drawn from a publicly available dataset comprising 2,149 patient records characterised by 34 demographic, lifestyle, clinical, and cognitive features. The proposed pipeline encompasses systematic data cleaning, categorical encoding, StandardScaler-based normalisation, and Synthetic Minority Oversampling Technique (SMOTE) application to address class imbalance. A novel consensus-driven multi-criteria feature selection strategy was developed, integrating Random Forest importance, Recursive Feature Elimination, and SHAP-based ranking into a unified normalised scoring mechanism to identify the 15 most diagnostically discriminative features. Eight classifiers spanning baseline and advanced architectural tiers were subsequently trained, optimised via stratified cross-validation, and evaluated across six performance metrics. LightGBM achieved the highest overall performance, attaining an accuracy of 94.88%, F1-Score of 92.62%, Cohen's Kappa of 0.8871, and AUCROC of 0.9491. Statistical validation through 10-fold cross-validation and Wilcoxon signed-rank testing confirmed the superiority of gradientboosting ensemble models over baseline classifiers at the $p < 0.05$ significance level. Based on explainability analysis, further clinically meaningful feature attribution, and enhanced model transparency and interpretability. The proposed framework demonstrates strong diagnostic potential and offers a scalable, cost-effective, and interpretable alternative to imaging-dependent AD diagnostic pipelines.

KEYWORDS: Alzheimer's disease; machine learning; explainable artificial intelligence; feature selection; LightGBM; SHAP; clinical decision support; ensemble learning.

How to Cite: Fatima Tauseef, Ahmad Jamal, Fahad Naseer, Md Shafayet Hossain, Dr. Mahzabin Mahmud, (2025) Towards Interpretable Alzheimer's Disease Diagnosis: A Multi-Model Machine Learning Framework with Consensus Feature Selection and SHAP-Driven Explainability, Vascular and Endovascular Review, Vol.8, No.2s, 345-357

INTRODUCTION

Alzheimer's disease (AD) is a chronic, progressive neurodegenerative disorder characterised by the gradual deterioration of memory, cognition, and functional independence, ultimately leading to severe disability and death[16]. It represents the most prevalent form of dementia globally, accounting for an estimated 60–70% of all dementia cases, and currently affects more than 55 million individuals worldwide, a figure projected to nearly triple by 2050 as populations age [17]. Beyond its devastating personal toll, AD imposes an extraordinary socioeconomic burden, with annual global costs exceeding one trillion US dollars and placing immense pressure on healthcare infrastructure, caregiving networks, and public health systems. Despite decades of intensive research, no diseasemodifying treatment has yet been approved, making early and accurate diagnosis the most viable strategy for slowing cognitive decline and improving patient quality of life [18].

Nevertheless, despite considerable scientific and clinical advances, accurately and efficiently diagnosing AD continues to pose significant challenges. Amyloid plaques and neurofibrillary tau tangles, among other pathological signs, begin accumulating in the brain without causing noticeable cognitive deficits for up to two to three decades before symptoms appear, opening a crucial period of diagnostic period that traditional approaches fail to capitalize on. Currently, AD diagnosis requires a series of subjective assessments and evaluations, expert knowledge on the part of the physician, and advanced, costly imaging techniques such as PET scans and high-resolution MRIs, which are not always accessible. Additionally, a wide array of overlapping cognitive symptoms complicates differentiating early AD, mild cognitive impairment (MCI), and natural aging, potentially leading to misdiagnoses that delay treatment at its most effective phase. Since there exists no definitive marker of AD, several clinical and imaging parameters have to be considered when diagnosing patients with early dementia, making the process complex, infrastructure-heavy, and lengthy. The aforementioned limitations call for innovative methods to reliably, objectively, and efficiently diagnose dementia. This creates a particular opportunity for AI and ML technologies, as they are capable of detecting intricate, multidimensional pathological patterns in neuroimaging, speech, and other data sets.

Recent AI approaches for diagnosing Alzheimer's disease have performed well on brain scans, speech, and health records, but there are still significant challenges before they can be widely used in clinics. Most studies focus on a single type of data, such as MRI or PET scans, and train their models on carefully selected datasets, such as ADNI. This makes it hard for these methods to work well with different groups of people or in everyday medical settings. While combining several types of data can improve accuracy, this often means relying on expensive scans, which isn't practical in places with limited resources. Using speech or language data is a less invasive option, but these methods often rely on small datasets that cover only certain languages and lack testing across diverse groups. Also, many models only classify the disease at a single point in time, instead of tracking how it progresses or predicting risks early on. Problems such as failing to test models on out-of-sample data, uneven datasets, and difficulty reproducing results are common. Plus, since many models are hard to interpret, doctors may hesitate to use them. All these issues point to a need for AI systems that can handle different types of data over time, work well with various populations, and provide clear explanations to support real clinical use.

To address the aforementioned gaps, this study proposes a comprehensive, clinically interpretable machine learning framework for automated classification of Alzheimer's disease using structured tabular clinical data, offering a scalable, cost-effective alternative to imaging-dependent diagnostic pipelines. The proposed framework integrates systematic preprocessing, SMOTE-based class balancing, and a novel consensus-driven multi-criteria feature selection strategy that unifies Random Forest importance, Recursive Feature Elimination, and SHAP-based ranking into a single normalised scoring mechanism. Eight classifiers spanning baseline and advanced architectural tiers are rigorously trained, optimised, and benchmarked under identical experimental conditions, with statistical significance testing and explainability analysis embedded throughout the evaluation pipeline. The principal contributions of this study are summarised as follows:

- A novel consensus-based feature selection framework combining three complementary ranking methods into a unified discriminative scoring mechanism, ensuring feature robustness across multiple evaluation criteria.
- A rigorous multi-model benchmarking pipeline incorporating stratified tenfold cross-validation and Wilcoxon signed-rank statistical testing to confirm the significance of reported performance differences.
- A comprehensive SHAP-driven explainability analysis providing both global feature attribution and individual prediction-level transparency, directly supporting clinical interpretability and trustworthiness.
- Empirical demonstration that gradient boosting ensemble models, particularly LightGBM, achieve state-of-the-art classification performance on structured clinical tabular data without reliance on costly neuroimaging modalities.

The paper is organized as follows: Section 2 presents the literature review, Section 3 describes the methodology, Section 4 reports the experimental results, and Section 5 concludes the study.

LITERATURE REVIEW

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the leading cause of dementia worldwide, making early and accurate diagnosis a pressing clinical priority. Conventional assessment methods often fail to detect subtle pathological changes in the earliest stages of the disease, prompting growing interest in artificial intelligence and machine learning as tools for automated, objective diagnosis. Recent research has pursued this goal through two principal avenues: deep learning models applied to neuroimaging data, such as structural MRI and PET, and non-invasive speech- and language-based biomarkers that offer scalable, cost-effective alternatives. The following review synthesizes key contributions across these domains, evaluating their methodological approaches, diagnostic performance, and limitations.

Recent advances in deep learning have substantially accelerated the development of automated systems for diagnosing Alzheimer's disease (AD), with a predominant focus on neuroimaging-based approaches. Li et al. (2024)[1] combined structural MRI (sMRI) and diffusion tensor imaging (DTI) mean diffusivity within a residual convolutional neural network framework, achieving an AD versus cognitively normal (CN) classification accuracy of 96.54% and an AUC of 0.9597, though the authors themselves acknowledged dataset size constraints and called for larger, more diverse cohorts. Building on single-modality architectures, El Assy et al. (2024)[2] proposed a dual-branch CNN operating exclusively on MRI that reported near-perfect accuracies of 99.43%, 99.57%, and 99.13% across three-, four-, and five-class classification settings, albeit with limited evidence of cross-site generalizability. To overcome the inherent information ceiling of unimodal imaging, Wang et al. (2024)[3] introduced an unsupervised cross-modal synthesis network that generates PET-like representations from MRI alone, demonstrating improved diagnostic performance on external AIBL and NACC datasets. Similarly, Tang et al. (2024)[5] proposed a dualbranch fusion CNN incorporating cross-modal enhancement and multi-scale attention gating, yielding exceptionally high accuracies of 99.59%, 98.85%, and 99.61% across pairwise diagnostic contrasts, while Odusami et al. (2024)[6] fused MRI with FDG-PET via Laplacian pyramid decomposition and a Mobile Vision Transformer, achieving precision values above 89% across all classification pairs. Complementing these multimodal strategies, Odimayo et al. (2024)[4] focused on structural atrophy in specific MRI sagittal views using a dedicated neural architecture with gamma-corrected preprocessing, achieving a sensitivity of 99.0% in mid-sagittal slices. The addition of explainability was addressed by Poonia et al. (2024)[7], whose ensemble of transfer learning models augmented with a Vision Transformer and explainable AI techniques achieved 96% accuracy on ADNI data, underscoring the growing attention to interpretability in clinical AI systems.

In parallel with neuroimaging-based methods, an emerging body of research has explored non-invasive, speech-based biomarkers as accessible alternatives for AD screening and progression prediction. Amini et al. (2024)[8] applied natural language processing and machine learning to neuropsychological interview speech combined with demographic features,

achieving 78.5% accuracy and 81.1% sensitivity in predicting MCI-to-AD conversion within a six-year window, though the study was constrained by a modest sample size of 166 participants and dependence on consistent recording conditions. Garc´ıa Guti´errez et al. (2024)[9] extended this line of inquiry to brief spontaneous speech tasks, employing paralinguistic and acoustic feature extraction to achieve F1 scores of 0.92 for AD and 0.84 for MCI, while also demonstrating meaningful correlations with multiple cognitive domains, suggesting that speech may carry broad neuropsychological signal beyond binary classification. Huang et al. (2024)[10] conducted a systematic comparison of acoustic, linguistic, and combined feature sets across multiple classifiers, finding that a support vector machine trained on all features yielded the highest accuracy of 80.77% for HC versus AD discrimination, though performance remained moderate for more clinically nuanced distinctions such as HC versus AD-or-MCI. Collectively, these studies reflect a field in rapid transition from single-modality imaging pipelines toward multimodal and multiformat paradigms, yet they also converge on shared limitations, including small or institutionally homogeneous datasets, reliance on ADNI-style cohorts, and insufficient external validation, all of which must be addressed before these systems can transition from research settings into routine clinical deployment. Table 1 summarizes recent studies on Alzheimer's disease prediction and monitoring, highlighting their methodologies, performance, and limitations.

METHODOLOGY

The methodology proposed in this study is structured as an end-to-end, clinically interpretable machine learning pipeline designed for the robust classification of Alzheimer's disease. As illustrated in Fig. 1, the framework encompasses six sequential stages—ranging from initial data acquisition and SMOTE-based balancing to consensus feature selection and multi-model evaluation. This systematic approach ensures that each phase of the diagnostic system is both analytically rigorous and fully reproducible.

Dataset Description

The Alzheimer's Disease dataset contains health-related information for 2,149 patients, with unique patient IDs ranging from 4751 to 6900 [19]. It includes a wide set of variables covering demographic characteristics, lifestyle habits, medical history, clinical measurements, cognitive and functional assessment scores, symptom indicators, and the final Alzheimer's disease diagnosis. The dataset exhibits moderate class imbalance, with 1,389 samples in class AD and 760 in class No AD. This dataset is suitable for statistical analysis, risk factor exploration, and machine learning-based prediction of Alzheimer's disease. Since it combines both clinical and behavioral information, it can support studies on diagnosis, feature importance, and the relationship between health conditions and cognitive decline. Table 2 summarizes the dataset features.

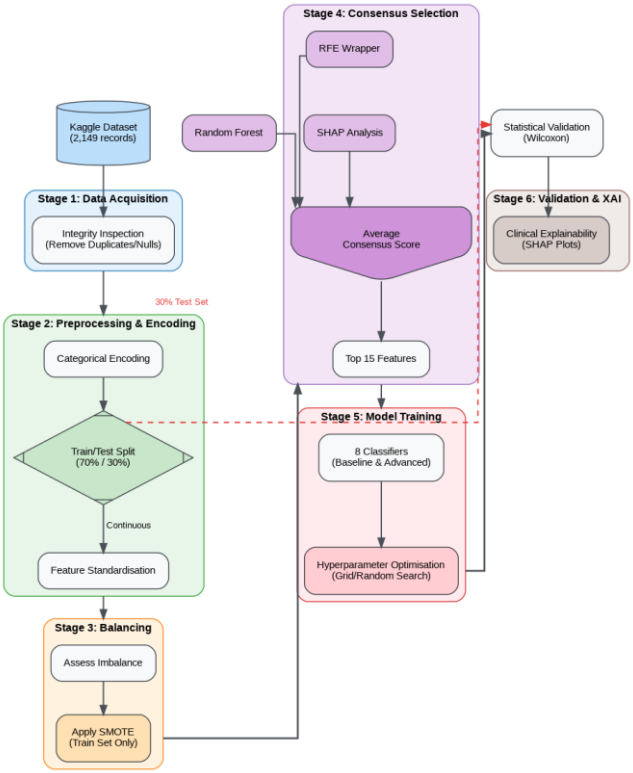


Fig.1: Proposed Framework Architecture Table 1: Summary of Recent Studies on Alzheimer's Disease Prediction and

Monitoring

Author	Contribution	Results	Limitations
Chou et al.[11]	Used linguistic features combined with biomarkers for AD prediction	Linguistic models achieved up to 88% accuracy; combined model reached AUC 0.93	Small sample size (80 participants); reliance on biomarker data such as PET and MRI
Sato et al.[12]	Estimated CDRSB performance scores mainly for MCI	For MCI (>= 18 months), R^2 >0.5	Strong for simplified and longitudinal monitoring longer follow-ups; not a full substitute for formal CDR	MAE <1.0, >0.5, MCC
Li et al.[13]	Early prediction of AD with real-world EHR data	Best AUCs: 0.939 (0 yr), 0.906 (1 yr), 0.884 (3 yr), 0.854 (5 yr)	Retrospective design; using performance decreases with longer prediction horizons
Fristed et al.[14]	Developed automated respeech-based screening for AD and amyloid positivity	AUC 0.77 (amyloid pos-itivity), 0.83 (MCI/mild AD)	Task-specific dataset; quires validation in realworld screening scenarios
Jahan et al.[15]	Proposed explainable multimodal framework for AD prediction and five-class classification	RF achieved 98.81% accuracy (10-fold CV)	Complex multimodal setup; requires validation on external datasets management

Exploratory Data Analysis

Exploratory Data Analysis was conducted to examine the distributions of major variables and their relationships with Alzheimer's disease diagnosis. Fig. 2a shows that patient ages are distributed across the range of 60 to 90 years, with a slight concentration toward the higher end. Similarly, Fig. 2b illustrates that BMI values span a wide range from approximately 15 to 40, without strong clustering in a specific interval. Overall, both distributions indicate sufficient variability in age and BMI, supporting robust predictive analysis.

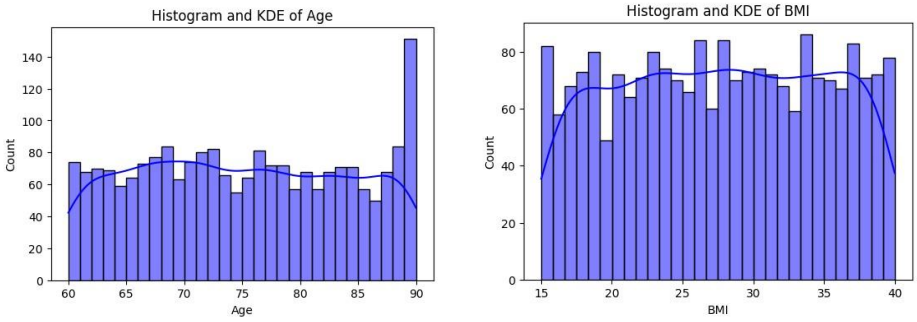
The relationship between MMSE and Age by Diagnosis, as shown in Fig. 3, shows a clear separation between diagnosed and non-diagnosed groups. Patients with Alzheimer's disease consistently exhibit lower MMSE scores than those without the disease across nearly all age values, indicating that cognitive impairment is strongly associated with the diagnosis variable.

In contrast, Fig. 4, the scatter plot of Alcohol Consumption and BMI, reveals no clear clustering or linear pattern between the two variables for either diagnosis group, suggesting that these factors alone may have limited discriminative value.
Fig. 5 cholesterol plots also show that total cholesterol, LDL, and HDL are widely scattered across age groups, with no strong age-driven pattern visible in the dataset.

Finally, the relationship between Education Level and MMSE suggests that education is only weakly associated with MMSE, although patients with a diTable 2: Description of Dataset Variables

Section	Variables	Description
Patient Information	PatientID	Unique identifier for each patient, ranging from 4751 to 6900
Demographic Details	Age, Gender, Ethnicity, EducationLevel	Includes age from 60 to 90 years, gender, ethnicity group, and education level
Lifestyle Factors	BMI, Smoking, AlcoholConsumption, PhysicalActivity, DietQuality, SleepQuality	Describes body condition, smoking status, alcohol intake, physical activity, diet score, and sleep quality
Medical History	FamilyHistoryAlzheimers, CardiovascularDisease, Diabetes, Depression, HeadInjury, Hypertension	Indicates whether the patient has relevant health history and risk-related conditions
Clinical Measurements	SystolicBP, DiastolicBP, CholesterolTotal, CholesterolLDL, CholesterolHDL, CholesterolTriglycerides	Includes blood pressure and cholesterol-related clinical values
Cognitive and Functional Assessments	MMSE, FunctionalAssessment, MemoryComplaints	Represents mental status, functional ability,

memory complaints, behav- ryComplaints, Symptoms entation, Personali-	Behav- ioral issues, and daily living capacity Confusion, Disori- Shows major symptoms commonly as- sociated with Alzheimer's disease tyChanges, DifficultyCompletingTasks, Forgetfulness		
Diagnosis	Informa- tion	Diagnosis	Target variable, where 0 indicates no Alzheimer's disease and 1 indicates Alzheimer's disease
Confidential	Infor- mation	DoctorInCharge	Contains confidential doctor-related information, with the same masked value for all records



(a) Age distribution of patients. (b) BMI distribution of patients.

Fig.2: Distribution of age and BMI among patients.

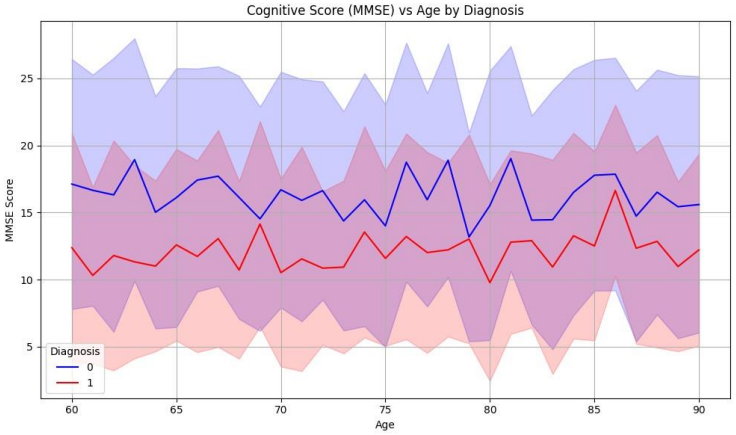


Fig.3: MMSE score across age by diagnosis group.

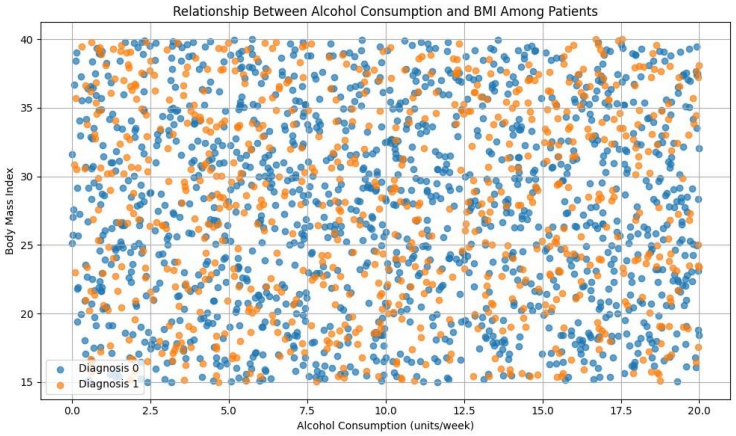


Fig.4: Alcohol consumption and BMI by diagnosis group.

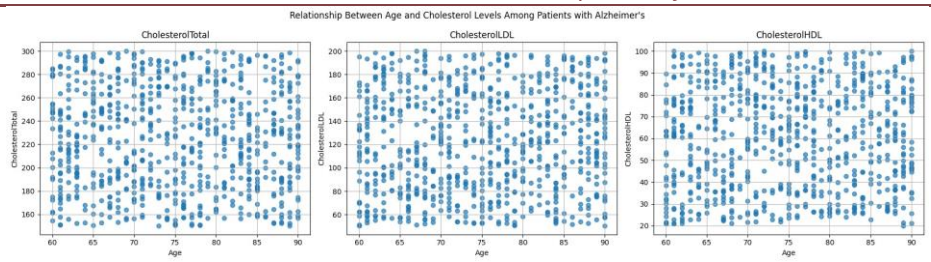


Fig.5: Age-wise distribution of cholesterol measures. agnosis tend to remain at lower MMSE levels across all education levels, as illustrated in Fig. 6.

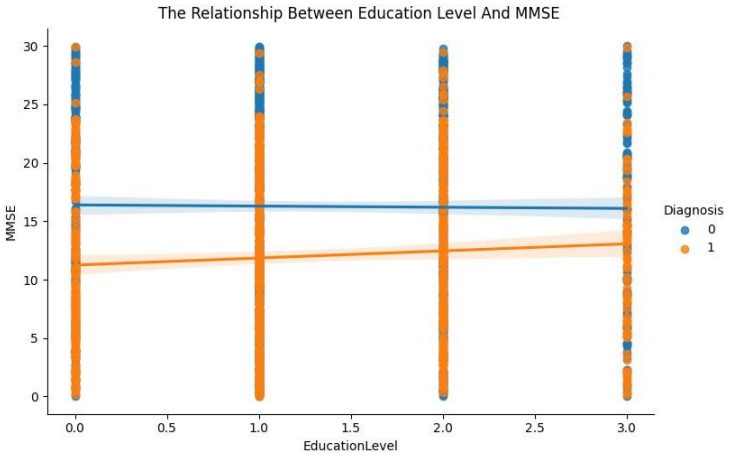


Fig.6: Relationship between education level and MMSE.

Overall, the EDA suggests that cognitive and functional variables are likely to be more informative predictors of Alzheimer's disease than general lifestyle or cholesterol-related variables.

Data Preprocessing

The dataset was systematically prepared to ensure quality and consistency for subsequent analysis. Duplicate records were identified and removed, missing values were handled appropriately, and a non-informative confidential dummy column was excluded from the feature set. Categorical variables, including Gender, Ethnicity, and Education Level, were encoded into numerical representations, and continuous features were assessed for variability, with high-range features standardised using a StandardScaler to prevent scale-induced bias during model training. The dataset was then partitioned into training (70%) and testing (30%) subsets using stratified random sampling to preserve the original class distribution across both partitions. Following this split, SMOTE was applied exclusively to the training set to generate synthetic minority class samples and address the observed class imbalance (760 AD versus 1,389 No AD), ensuring that the test set remained entirely untouched and free from any synthetic observations throughout the balancing procedure. The StandardScaler was subsequently fitted solely on the training data and applied to both partitions, preventing any leakage of test set information into the preprocessing pipeline. The complete preprocessing pipeline is illustrated in Fig 7.

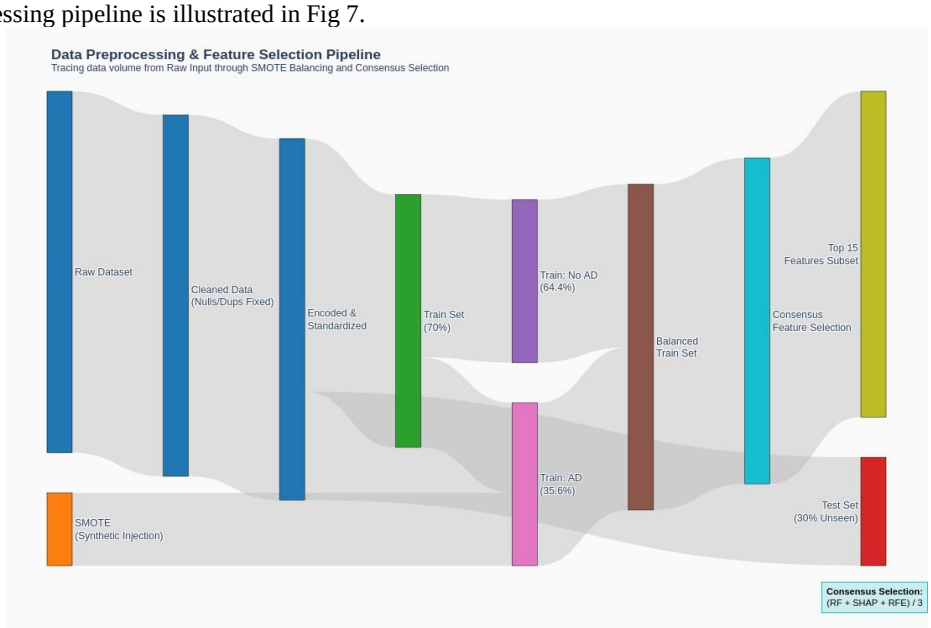


Fig.7: Data Preprocessing Pipeline

Feature Selection

A consensus-based multi-criteria feature selection framework was applied to identify the most diagnostically relevant features and reduce dimensionality, as illustrated in Fig. 8. Three complementary methods were employed. First, a Random Forest classifier was used to compute global feature importance based on impurity reduction. Second, Recursive Feature Elimination (RFE) was applied to rank features by iteratively removing the least important ones, retaining the top 15. Third, SHAP values were calculated using a TreeExplainer, with the mean absolute SHAP value as the importance measure.

To obtain a unified ranking, the importance scores from all three methods were normalized to the range [0, 1] using Min-Max scaling and then averaged to compute a final consensus score for each feature, as defined in Equation (1). The top 15 features, ranked by this score, were selected for model training and evaluation.

$$\text{Final Score}_i = \frac{\text{RF norm}_i + \text{SHAP norm}_i + \text{RFE norm}_i}{3} \quad (1)$$

This approach ensures that selected features are consistently important across multiple criteria, improving robustness and interpretability.

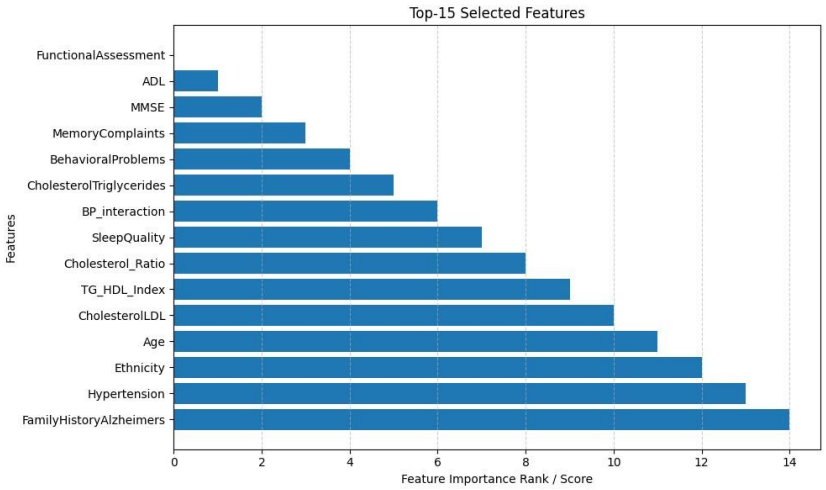


Fig.8: Top features selected using the proposed consensus-based method.

Classification Models

In this study, eight machine learning classifiers were employed to evaluate the diagnostic classification of Alzheimer's disease, spanning from interpretable linear baselines to advanced ensemble and deep learning architectures. This multimodel strategy was deliberately adopted to enable a rigorous comparative analysis and to contextualize the performance of advanced models against wellestablished benchmarks under identical experimental conditions.

Logistic Regression (LR) was employed as the primary linear baseline, modeling the log-odds of the binary diagnosis outcome as a linear combination of input features. Despite its simplicity, LR provides a highly interpretable probabilistic output and serves as an essential reference point for evaluating the added value of more complex classifiers.

K-Nearest Neighbors (KNN) is a non-parametric instance-based classifier that assigns a class label based on the majority vote among the K closest training samples in feature space. With K set to 5, it captures local data structure without assuming any underlying distributional form, making it a useful complement to model-based approaches.

Support Vector Machine (SVM) was applied with a Radial Basis Function (RBF) kernel, which projects the input features into a higher-dimensional space to identify an optimal separating hyperplane with maximum margin between the AD and No AD classes. Probability calibration was enabled to facilitate AUC-ROC computation and threshold-independent evaluation.

Random Forest (RF) is a bagging ensemble method that constructs a large number of decorrelated decision trees during training and aggregates their predictions through majority voting. With 200 estimators, RF provides strong generalization, inherent resistance to overfitting, and native feature importance estimation, making it particularly well-suited to high-dimensional clinical tabular data.

Gradient Boosting (GB) builds an ensemble of decision trees sequentially, where each successive tree is trained to correct the residual errors of its predecessor. This iterative refinement process enables the model to capture complex non-linear feature interactions, and 200 estimators were used to ensure sufficient learning capacity.

Extreme Gradient Boosting (XGBoost) extends the standard gradient boosting framework of incorporating L1 and L2 regularization, parallel tree construction, and efficient handling of sparse data. These enhancements confer improved predictive performance and robustness over traditional gradient boosting, and the model was configured with 200 estimators and log-loss as the evaluation metric.

Light Gradient Boosting Machine (LightGBM) employs a histogrambased leaf-wise tree growth strategy that substantially reduces memory consumption and training time relative to conventional level-wise boosting algorithms. Its efficiency and strong performance on structured tabular datasets make it particularly appropriate for clinical classification tasks, and it was trained with 200 estimators.

Multilayer Perceptron (MLP) is a feedforward artificial neural network comprising three fully connected hidden layers with 128, 64, and 32 neurons respectively, using non-linear activation functions to capture highly complex feature interactions that linear and tree-based models may fail to model. The network was trained for a maximum of 500 iterations using backpropagation with adaptive learning rate optimization.

All models were trained exclusively on the SMOTE-balanced training set and evaluated on the untouched holdout test set. Models sensitive to feature scale, specifically LR, KNN, SVM, and MLP, received standardized input features processed using a StandardScaler fitted solely on the training data, ensuring no data leakage into the evaluation pipeline.

Hyperparameter Optimization

Table 3 presents the optimal hyperparameter configurations for all eight classification models used in this study. It summarizes the best-performing settings obtained through systematic hyperparameter tuning under a consistent crossvalidation framework. This allows a fair and reproducible comparison of all models under their respective optimized conditions.

RESULT AND DISCUSSION

Classification Performance of Individual Models

The evaluation of the eight machine learning models reveals a clear performance advantage for advanced tree-based ensemble architectures over traditional baseline algorithms. As detailed in Table 4, LightGBM, Random Forest, and XGBoost emerged as the top-performing classifiers across almost all evaluation Table 3: Hyperparameter configurations and tuning results for all eight classification models. CV = cross-validation; LR = learning rate.

Model	Tier	Hyperparameter space	Search method	Optimal configuration
Baseline models				
Logistic Regression	Baseline	C∈{0.001, 0.01, 0.1, 1, 10, 100}; solver∈ {lbfgs, liblinear}; max_iter∈{500, 1000}	Grid Search + 10-fold CV	C=1, solver=lbfgs, max_iter=1000
K-Nearest Neighbors	Baseline	n_neighbors∈{3, 5, 7, 9, 11}; weights∈ {uniform, distance}; metric∈{euclidean, manhattan}	Grid Search + 10-fold CV	k=5, weights=uniform, metric=euclidean
Support Vector Machine	Baseline	C∈{0.1, 1, 10, 100}; kernel∈{rbf, poly, linear}; gamma∈{scale, auto}	Grid Search + 10-fold CV	C=10, kernel=rbf, gamma=scale
Advanced models				
Random Forest	Advanced	n_estimators ∈ {100, 200, 300}; max_depth ∈ {None, 10, 20, 30}; min_samples_split ∈ {2, 5, 10}; max_features∈{sqrt, log2}	Random Search + 10-fold CV	n_estimators=200, max_depth=None, min_samples_split=2, max_features=sqrt
Gradient Boosting	Advanced	n_estimators ∈ {100, 200, 300}; learning_rate ∈ {0.01, 0.05, 0.1, 0.2}; max_depth ∈ {3, 5, 7}; subsample ∈ {0.7, 0.8, 1.0}	Random Search + 10-fold CV	n_estimators=200, lr=0.1, max_depth=5
XGBoost	Advanced	n_estimators ∈ {100, 200, 300}; learning_rate ∈ {0.01, 0.05, 0.1}; max_depth ∈ {3, 6, 9}; reg_alpha ∈ {0, 0.1, 1}; reg_lambda∈{1, 1.5, 2}	Random Search + 10-fold CV	n_estimators=200, eval_metric=logloss lr=0.1, max_depth=6,
LightGBM	Advanced	n_estimators ∈ {100, 200, 300}; learning_rate ∈ {0.01, 0.05, 0.1}; num_leaves ∈ {31,63,127}; min_child_samples∈{10, 20, 30}	Random Search + 10-fold CV	n_estimators=200, lr=0.05, num_leaves=63
MLP Neural Network	Advanced	hidden_layer_sizes ∈ {(64,32), (128,64,32), (256,128,64)}; activation ∈ {relu, tanh}; learning_rate_init ∈ {0.001, 0.0001}; max_iter∈{300, 500}	Grid Search + 10-fold CV	hidden_layer_sizes=(128,64,32), activation=relu, lr=0.001, max_iter=500

metrics. LightGBM achieved the highest overall accuracy (94.88%), precision (94.52%), F1-score (0.9262), and Cohen's Kappa (0.8871), indicating an exceptional balance between correctly identifying Alzheimer's cases and minimizing false positives.

Furthermore, the ensemble models demonstrated outstanding discriminatory ability, with Area Under the Curve (AUC) scores consistently exceeding 0.94, peaking at 0.9504 for the Random Forest classifier.

In contrast, the baseline models and the Multilayer Perceptron (MLP) yielded moderate diagnostic performance, albeit lower than that of the other models. Support Vector Machine (SVM) and the MLP Neural Network were the strongest among the non-ensemble methods, achieving accuracies of 85.58% and 86.20%, respectively. However, algorithms like K-Nearest Neighbors (KNN) and Logistic Regression struggled to match the predictive power of the advanced models, recording the lowest overall F1-scores and Kappa statistics. This performance disparity highlights the ability of robust, nonlinear ensemble frameworks to more effectively capture the complex clinical and demographic patterns in the Alzheimer's disease dataset.

Table 4: Classification performance of individual machine learning models on the holdout test set. Best results in each column are highlighted in bold.

Model	ACC	PRE	REC	F1	KAPPA	AUC
Logistic Regression	0.8357	0.7723	0.7588	0.7655	0.6390	0.8966
K-Nearest Neighbors	0.8047	0.7237	0.7237	0.7237	0.5726	0.8599
Support Vector Machine	0.8558	0.8169	0.7632	0.7891	0.6798	0.9121
Random Forest	0.9457	0.9447	0.8991	0.9213	0.8800	**0.9504**
Gradient Boosting	0.9395	0.9127	**0.9167**	0.9147	0.8678	0.9459
XGBoost	0.9457	0.9367	0.9079	0.9220	0.8804	0.9482
LightGBM	**0.9488**	**0.9452**	0.9079	**0.9262**	**0.8871**	0.9491
MLP Neural Network	0.8620	0.8145	0.7895	0.8018	0.6960	0.9060

ROC and Precision-Recall Analysis

Fig. 9 and Fig. 10 present the ROC and Precision Recall curves for all classification models. The ROC analysis shows that the tree-based ensemble models achieved the strongest discrimination performance. Random Forest achieved the highest AUC of 0.9504, followed closely by XGBoost (0.9482), Gradient Boosting (0.9459), and LightGBM (0.9491)rast, K Nearest Neighbor K-Nearest Neighbors produced the lowest ROC performance,8599. These results indicate that ensemble learning methods were more effective in separating Alzheimer's and non-Alzheimer's cases than the baseline models.

The Precision Recall analysis further confirms the superiority of the advanced ensemble models. LightGBM achieved the highest average precision score of 0.9271, followed by Gradient Boosting with 0.9232, XGBoost with 0.9196, and Random Forest with 0.9186. Since Precision Recall curves are more informative for medical classification tasks where false predictions are important, these findings suggest that the boosting-based models offered the most reliable predictive behavior. Overall, the results show that ensemble methods provided stronger and more stable classification performance than Logistic Regression, K Nearest Neighbors, Support Vector Machine, and the MLP Neural Network.

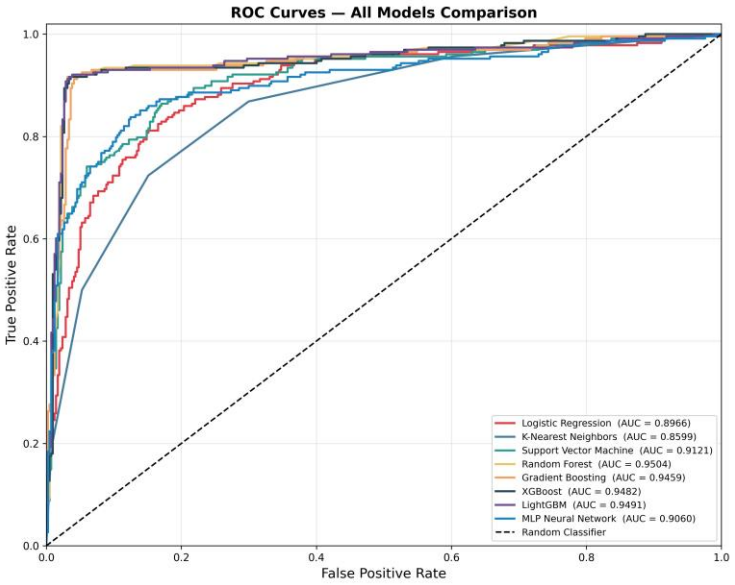


Fig.9: ROC Curve Comparison of All Models

Statistical Robustness and Cross-Validation Analysis

To assess the stability and reliability of the developed classifiers, a stratified ten-fold cross-validation analysis was conducted using AUC-ROC as the evaluation metric, with LightGBM designated as the reference model for all pairwise Wilcoxon signed-rank comparisons given its highest observed mean AUC-ROC across all folds. The results presented in Table 5 demonstrate that

ensemblebased methods achieved the strongest and most consistent performance across folds. LightGBM produced the highest mean AUC-ROC of 0.9560±0.0171, followed closely by Random Forest (0.9546 ± 0.0167), XGBoost (0.9535 ± 0.0174), and Gradient Boosting (0.9518 ± 0.0205). These results confirm that advanced ensemble models maintained both high discrimination ability and low cross-fold variance, indicating strong generalisation capacity.

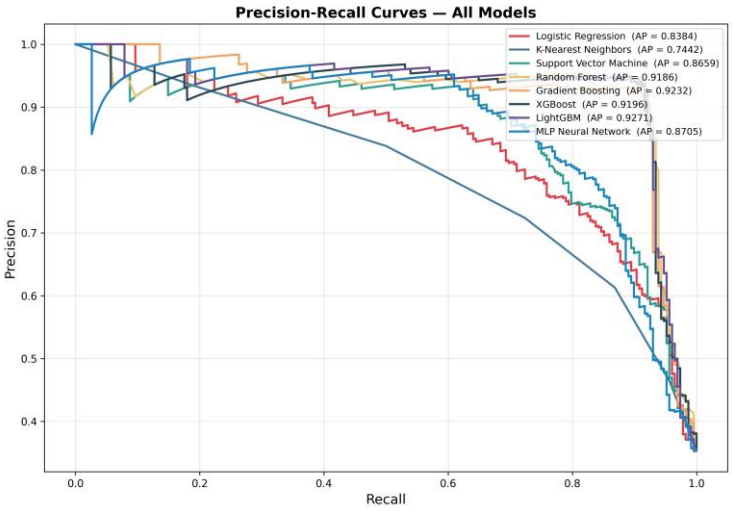


Fig.10: Precision Recall Curve Comparison of All Models

In contrast, baseline models demonstrated notably weaker performance. Logistic Regression achieved a mean AUC-ROC of 0.9051±0.0268, Support Vector Machine achieved 0.9231 ± 0.0247, and K-Nearest Neighbors produced the lowest mean AUC-ROC of 0.8627 ± 0.0260. The MLP Neural Network achieved 0.9001 ± 0.0368 but exhibited the largest standard deviation among all models, indicating comparatively unstable behavior during repeated validation on this tabular dataset.

Wilcoxon signed-rank tests confirmed that LightGBM significantly outperformed all three baseline classifiers and the MLP Neural Network at the $p < 0.05$ threshold. No statistically significant difference was observed between LightGBM and the remaining advanced ensemble models — Gradient Boosting ($p = 1.0000$), Random Forest ($p = 0.6250$), and XGBoost ($p = 0.6953$) — indicating that these three gradient boosting variants converge to a statistically comparable performance ceiling on this dataset rather than one being definitively superior to the others.

The boxplot in Fig. 11 further supports these findings by showing that the advanced ensemble models have both higher median AUC values and tighter interquartile ranges than the conventional models. Although LightGBM achieved the highest average AUC, the Wilcoxon signed-rank test shows that the performance differences among the strongest ensemble methods were not statistically significant compared with the selected reference best model. In particular, the differences between Gradient Boosting ($p = 1.0000$), XGBoost ($p = 0.6250$), and LightGBM ($p = 0.6953$) were not significant, indicating that these topperforming methods are statistically comparable. On the other hand, Logistic Regression ($p = 0.0020$), K Nearest Neighbors ($p = 0.0020$), Support Vector Machine ($p = 0.0039$), and MLP Neural Network ($p = 0.0020$) showed statisTable 5: Ten-fold cross-validation AUC-ROC results and Wilcoxon signed-rank test comparisons against the best-performing model (LightGBM). Significance threshold: $p < 0.05$.

Model	**Tier**	**CV AUC-ROC (Mean ± Std)**	***p*-value**	**Significance**
Baseline models				
Logistic Regression	Baseline	0.9051 ± 0.0268	0.0020	Significant*
K-Nearest Neighbors	Baseline	0.8627 ± 0.0260	0.0020	Significant*
Support Vector Machine	Baseline	0.9231 ± 0.0247	0.0039	Significant*
Advanced models				
Gradient Boosting	Advanced	0.9518 ± 0.0205	1.0000	Not significant
Random Forest	Advanced	0.9546 ± 0.0167	0.6250	Not significant
XGBoost	Advanced	0.9535 ± 0.0174	0.6953	Not significant
LightGBM	Advanced	**0.9560 ± 0.0171**	—	**Best model (Reference)**

MLP Network	Neural Advanced	0.9001 ± 0.0368	0.0020 Significant*

*

Statistically significant difference from LightGBM ($p < 0.05$, two-tailed Wilcoxon signed-rank test).

LightGBM serves as the reference model for all pairwise comparisons , given its highest mean AUC-ROC across ten folds. A dash (—) in the *p*-value column indicates the reference model itself, for which self-comparison is not applicable. Gradient Boosting, Random Forest and XGBoost showed no statistically significant difference from LightGBM, indicating comparable performance within the advanced ensemble tier.

tically significant differences, confirming that their predictive performance was meaningfully lower. Overall, the cross validation and significance testing results demonstrate that ensemble learning methods provide the most robust and reliable framework for Alzheimer's disease prediction in this study.

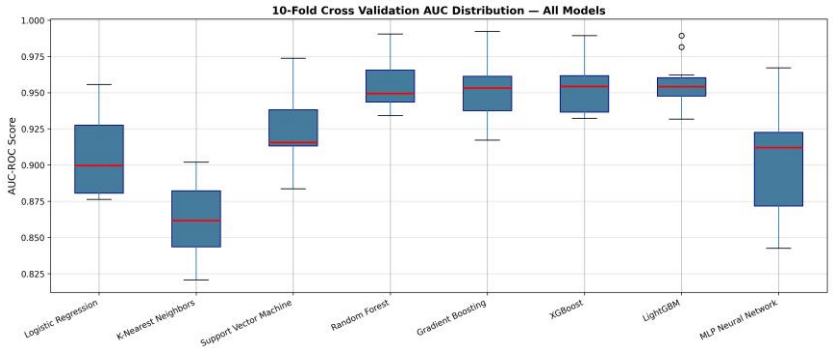


Fig.11: Distribution of 10-fold cross-validation AUC scores across all classification models.

These findings collectively indicate that the advanced gradient boosting ensemble models, particularly LightGBM, XGBoost, and Gradient Boosting, demonstrated not only superior predictive accuracy but also greater cross-validation stability compared to baseline classifiers and the MLP Neural Network, the latter of which exhibited the highest performance variance across folds. Whilst statistically significant performance differences were confirmed between LightGBM and all three baseline models as well as the MLP (p¡0.05p ¡ 0.05 p¡0.05), no statistically significant difference was observed among the top three gradient boosting variants, suggesting that these models converge to a comparable performance ceiling on this tabular clinical dataset rather than one being definitively superior to the others.

Calibration Analysis

The calibration curves show how closely each model's predicted probabilities match the actual observed outcomes. Models whose curves stay nearer to the diagonal reference line are better calibrated, meaning their probability estimates are more reliable. In Fig. 12, Logistic Regression and Support Vector Machine follow the reference line more closely over several probability ranges, indicating relatively stable calibration. In contrast, K-Nearest Neighbors and the MLP Neural Network exhibit larger deviations from the diagonal, suggesting less consistent probability estimates. The ensemble models, especially Random Forest, Gradient Boosting, XGBoost, and LightGBM, achieve strong classification performance, but their calibration curves deviate in some regions, which means their predicted probabilities may be slightly overconfident or underconfident at certain levels. Overall, the figure suggests that some models are strong discriminators, while others provide more trustworthy probability estimates.

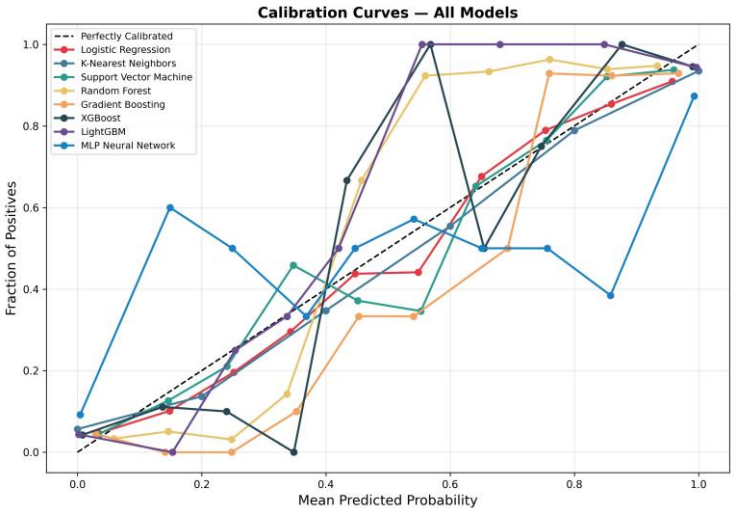


Fig.12: Calibration curves of all classification models for probability estimation analysis.

SHAP Analysis

The SHAP waterfall plot shows how individual features influenced the prediction for a single patient, as shown in Fig. 13. The model starts with the average prediction level, $E[f(X)] = 0.314$, and then adjusts this value based of each feature's contribution, ultimately reaching $f(x) = 0.015$. In this case, ADL and FunctionalAssessment made the strongest negative contributions, each reducing the prediction by about 0.10. The absence of MemoryComplaints and BehavioralProblems also lowered the prediction, while MMSE made a small positive contribution. This means that, for this patient, the combined effect of functional ability and the absence of major symptoms outweighed the small upward influence of MMSE, leading the model to assign a very low probability to the positive class.

The SHAP dependence plot shows how the value of FunctionalAssessment affects the model's prediction across all patients, as shown in Fig. 14. Lower FunctionalAssessment values are mostly associated with positive SHAP values, which means they increase the predicted likelihood of Alzheimer's disease. In contrast, higher FunctionalAssessment values are mostly associated with negative SHAP values, indicating they reduce the prediction of the disease class. A clear shift appears around the middle of the scale, where the SHAP contribution changes from positive to negative. The color coding by ADL suggests that the effect of FunctionalAssessment is not isolated, but interacts with daily living ability. This indicates that functional impairment plays an important role in the model, and its impact becomes stronger when combined with related daily living characteristics.

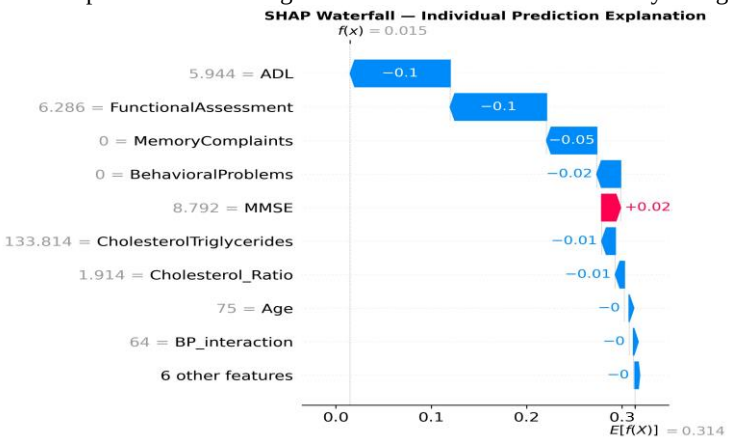


Fig.13: SHAP waterfall explanation for an individual patient prediction.

CONCLUSION

This study presented a comprehensive, clinically interpretable machine learning framework for automated classification of Alzheimer's disease using structured tabular clinical data. The proposed pipeline integrates systematic data preprocessing, SMOTE-based class balancing, and a novel consensus-driven multicriteria feature selection strategy combining Random Forest importance, Recursive Feature Elimination, and SHAP-based ranking, followed by rigorous training and evaluation of eight classifiers spanning baseline and advanced architectural tiers. Experimental results demonstrated that gradient boosting ensemble models, particularly LightGBM, XGBoost, and Random Forest, consistently outperformed baseline classifiers across all evaluation metrics, with LightGBM achieving the highest overall performance of 94.88% accuracy, 92.62% F1-Score, and an AUC-ROC of 0.9491. Statistical validation via ten-fold cross-validation and Wilcoxon signed-rank testing confirmed that these performance advantages were

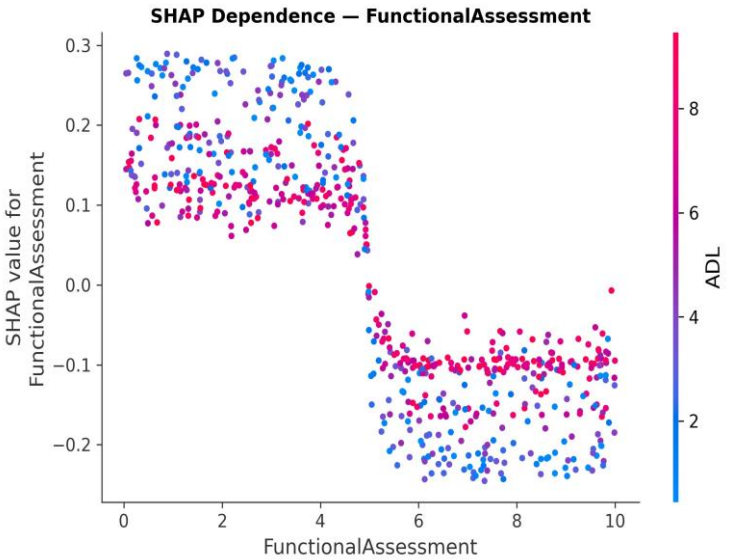


Fig.14: SHAP dependence plot showing the effect of FunctionalAssessment on model prediction.

statistically significant rather than incidental, reinforcing the reliability and generalisability of the proposed framework. Furthermore, the integration of SHAPbased explainability provided clinically meaningful feature attribution, identifying the most diagnostically influential variables and enhancing the transparency and trustworthiness of model predictions for potential clinical adoption. Despite these contributions, several limitations merit acknowledgment. The framework was developed and evaluated on a single publicly available dataset, and external validation on independent, prospectively collected clinical cohorts remains necessary to confirm real-world generalisability. Additionally, the binary classification scope does not address the nuanced staging of MCI or disease progression trajectories. Future work will pursue multimodal data integration incorporating neuroimaging and biomarker features, longitudinal modeling to predict disease progression from MCI to AD, and federated learning approaches to enable privacy-preserving multi-institutional validation, collectively advancing the framework toward deployment-ready clinical decision support.

REFERENCES

1. Li, Y., Chen, G., Wang, G., Zhou, Z., An, S., Dai, S., ... & Yu, F. (2024). Dominating Alzheimer's disease diagnosis with deep learning on sMRI and DTI-MD. Frontiers in neurology, 15, 1444795.
2. El-Assy, A. M., Amer, H. M., Ibrahim, H. M., & Mohamed, M. A. (2024). A novelCNN architecture for accurate early detection and classification of Alzheimer's disease using MRI data. Scientific Reports, 14(1), 3463.
3. Wang, C., Piao, S., Huang, Z., Gao, Q., Zhang, J., Li, Y., ... & Alzheimer's DiseaseNeuroimaging Initiative. (2024). Joint learning framework of cross-modal synthesis and diagnosis for Alzheimer's disease by mining underlying shared modality information. Medical Image Analysis, 91, 103032.
4. Odimayo, S., Olisah, C. C., & Mohammed, K. (2024). SNeurodCNN: Structurefocused Neurodegeneration Convolutional Neural Network for Modelling and Classification of Alzheimer's Disease. arXiv preprint arXiv:2401.03922.
5. Tang, C., Xi, M., Sun, J., Wang, S., Zhang, Y., & Alzheimer's Disease Neuroimaging Initiative. (2024). MACFNet: Detection of Alzheimer's disease via multiscale attention and cross-enhancement fusion network. Computer Methods and Programs in Biomedicine, 254, 108259.
6. Odusami, M., Damaˇseviˇcius, R., Milieˇskaiteˇ-Belousovieneˇ, E., & Maskeliu¯nas, R.(2024). Alzheimer's disease stage recognition from MRI and PET imaging data using Pareto-optimal quantum dynamic optimization. Heliyon, 10(15).
7. Poonia, R. C., & Al-Alshaikh, H. A. (2024). Ensemble approach of transfer learning and vision transformer leveraging explainable AI for disease diagnosis: An advancement towards smart healthcare 5.0. Computers in Biology and Medicine, 179, 108874.
8. Amini, S., Hao, B., Yang, J., Karjadi, C., Kolachalama, V. B., Au, R., & Paschalidis, I. C. (2024). Prediction of Alzheimer's disease progression within 6 years using speech: A novel approach leveraging language models. Alzheimer's & Dementia, 20(8), 5262-5270.
9. Garc´ıa-Guti´errez, F., Alegret, M., Marqui´e, M., Munˇoz, N., Ortega, G., Cano,A., ... & Valero, S. (2024). Unveiling the sound of the cognitive status: Machine Learning-based speech analysis in the Alzheimer's disease spectrum. Alzheimer's Research & Therapy, 16(1), 26.
10. Huang, L., Yang, H., Che, Y., & Yang, J. (2024). Automatic speech analysis fordetecting cognitive decline of older adults. Frontiers in Public Health, 12, 1417966.
11. Chou, C. J., Chang, C. T., Chang, Y. N., Lee, C. Y., Chuang, Y. F., Chiu, Y.L., ... & Liu, Y. C. (2024). Screening for early Alzheimer's disease: enhancing diagnosis with linguistic features and biomarkers. Frontiers in Aging Neuroscience, 16, 1451326.
12. Sato, K., Niimi, Y., Ihara, R., Suzuki, K., Iwata, A., Iwatsubo, T., ... & JapaneseAlzheimer's Disease Neuroimaging Initiative. (2024). Simplifying Alzheimer's Disease Monitoring: A Novel Machine-Learning Approach to Estimate the Clinical Dementia Rating Sum of Box Changes Using the Mini-Mental State Examination and Functional Activities Questionnaire. Journal of Alzheimer's Disease, 99(3), 953-963.
13. Li, Q., Yang, X., Xu, J., Guo, Y., He, X., Hu, H., ... & Bian, J. (2023). Earlyprediction of Alzheimer's disease and related dementias using real-world electronic health records. Alzheimer's & Dementia, 19(8), 3506-3518.
14. Fristed, E., Skirrow, C., Meszaros, M., Lenain, R., Meepegama, U., Papp, K. V.,... & Weston, J. (2022). Leveraging speech and artificial intelligence to screen for early Alzheimer's disease and amyloid beta positivity. Brain Communications, 4(5), fcac231.
15. Jahan, S., Abu Taher, K., Kaiser, M. S., Mahmud, M., Rahman, M. S., Hosen, A. S.,& Ra, I. H. (2023). Explainable AI-based Alzheimer's prediction and management using multimodal data. Plos one, 18(11), e0294253.
16. Knopman, D. S., Amieva, H., Petersen, R. C., Ch´etelat, G., Holtzman, D. M.,Hyman, B. T., ... & Jones, D. T. (2021). Alzheimer disease. Nature reviews Disease primers, 7(1), 33.
17. World Health Organization. (2024, March 31). Dementia [Fact sheet]. https://www.who.int/news-room/fact-sheets/detail/dementia
18. Wimo, A., Seeher, K., Cataldi, R., Cyhlarova, E., Dielemann, J. L., Frisell, O., ... &Dua, T. (2023). The worldwide costs of dementia in 2019. Alzheimer's & dementia, 19(7), 2865-2873.
19. El Kharoua, R. (2024). Alzheimer's disease dataset [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/8668279