Cross-Model Interpretability of Alzheimer's Disease Classification Using ADNI Dataset

Ghada Abdulsalam*, Moataz Ahmed*†, Mufti Mahmud*†‡
* ICS Department, King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia
† SDAIA-KFUPM JRC for AI, King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia
‡ IRC for Bio Systems and Machines, King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia
Emails: {g202212880, moataz, mufti.mahmud}@kfupm.edu.sa, muftimahmud@gmail.com

***Abstract*—Alzheimer's Disease (AD) staging from tabular clinical and neuropsychological data remains highly relevant for real-world decision support, yet model interpretability is hindered by multicollinearity among overlapping cognitive and functional measures. Using baseline Alzheimer's Disease Neuroimaging Initiative (ADNI) data, we study one-vs-rest classification of Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and AD with linear and tree-based models: Elastic Net Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). Beyond predictive performance, we provide a structured interpretability workflow that integrates Top-K global importance, cross-model feature agreement, correlation analysis of consensus predictors, and SHapley Additive exPlanations (SHAP). Results show a stable, clinically coherent core of predictors across models, with stage-specific shifts from global cognition and memory signals in CN to combined cognitive–functional impairment in MCI, and to severity and activities-of-daily-living measures in AD. Cross-model agreement highlights clinical dementia rating and functional measures as robust AD indicators. SHAP summaries further demonstrate consistent directional effects for key measures (e.g., higher CDRSB increasing the likelihood of AD and higher MMSE decreasing it). Overall, the proposed framework improves interpretability by contextualizing feature importance with cross-model robustness and correlation structure, supporting more reliable identification of AD-relevant predictors from tabular data.**

***Index Terms*—Alzheimer's disease classification, Explainable artificial intelligence, Feature importance, SHAP, ADNI dataset, Cross-model interpretability**

I. Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder and the leading cause of dementia worldwide. Accurate discrimination among Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and AD stages is crucial for early diagnosis and clinical decision-making, motivating the extensive use of machine learning (ML) with heterogeneous data, including neuroimaging and cognitive assessments [1], [2]. Large-scale datasets, notably ADNI, support systematic evaluation of ML models using standardised clinical and neuropsychological features [3]. While deep learning excels in imaging tasks, tabular clinical data remain highly informative and offer advantages in robustness and interpretability [4].

Despite progress, interpretability limits clinical adoption. Feature importance is often derived from a single model or metric, neglecting multicollinearity and cross-model variability common in ADNI features, which can yield unstable or misleading explanations [5], [6]. Explainable AI methods, particularly SHAP, provide unified local and global explanations for linear and tree-based models [7], [8], but existing AD studies rarely assess robustness across models or relate explanations to feature correlation structure [9].

In this work, we study AD classification using Elastic Net Logistic Regression, Random Forest, and XGBoost on ADNI-derived clinical and cognitive features. We propose a structured interpretability framework combining global importance, cross-model agreement, correlation analysis, and SHAP explanations to identify robust and clinically meaningful predictors. The remainder of the paper is organised as follows: Section II reviews related work, Section III describes the methodology, Section IV presents the interpretability framework, Section V discusses the results, and Section VI concludes the paper.

II. Related Work

Machine learning (ML) approaches for AD diagnosis and staging have been widely studied using neuroimaging, clinical, and neuropsychological data. The methodologies range from foundational deep learning surveys [10] to high-precision ensemble methods, including 3D DenseNets [11] and Siamese networks for 4-way classification [12], [13], [14]. Research has further expanded into early detection from clinical data [15], [16], local field potentials [17], IoHT-based care frameworks [18] and securing these frameworks [19], [20], [21], and multimodal prediction models [22], [23]. Currently, a significant paradigm shift toward Explainable AI (XAI) emphasises clinical transparency, leveraging LIME and SHAP to ensure the interpretability of diagnostic models [24], [7]. Also, studies show that cognitive and functional assessments from datasets such as ADNI remain informative for discriminating CN, MCI, and AD, particularly in tabular and multimodal settings [25], [26].

Logistic Regression (LR) and its regularised variants remain strong and interpretable baselines for AD classification. Elastic Net regularisation is particularly effective in high-dimensional clinical settings, balancing sparsity and stability in the presence of correlated predictors [27]. ADNI-based studies frequently use penalised regression to identify clinically meaningful predictors and transparent decision boundaries [28], [29]. However, LR is sensitive to multicollinearity, often

producing unstable or distributed importance estimates when cognitive and functional measures are highly correlated.

Tree-based ensemble methods, especially Random Forest (RF) and XGBoost, are among the most frequently reported high-performing models for AD classification with heterogeneous clinical data. RF is robust to noise and correlated features [30], while XGBoost captures nonlinear interactions through gradient-boosted trees [31]. Although these models often outperform linear baselines, their feature-importance estimates can vary substantially across model structures and definitions of importance, complicating interpretation [32], [33].

Feature selection and importance analysis are therefore central to AD detection, particularly in high-dimensional data. Prior work has explored filter-, wrapper-, and embedded-based approaches [34], [35], but many studies rely on a single model or importance metric, without assessing robustness across model families or accounting for feature collinearity. As a result, reported important features may reflect model-specific artefacts rather than stable disease-related signals.

XAI methods, notably SHAP, have gained attention for providing unified explanations for both linear and tree-based models [36], [37]. However, recent studies show that SHAP explanations remain model-dependent, reinforcing the need for cross-model validation [38]. In contrast to prior work, our study jointly examines global importance, cross-model agreement, correlation structure, and SHAP-based explanations to identify robust and clinically meaningful predictors of AD.

III. Methodology

A. Dataset and Feature Preparation

We utilise baseline data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [39], a large, multi-centre longitudinal study designed to develop biomarkers for the early detection and tracking of disease progression in Alzheimer's Disease. The dataset includes demographic variables, clinical assessments, and neuropsychological scores. Diagnostic labels include CN, MCI, and AD.

For data preprocessing, the missing values were imputed using the median computed on the training data only, followed by standardisation to zero mean and unit variance:

$$x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j},$$

where μ_j and σ_j denote the mean and standard deviation of feature j in the training set. The multi-class diagnosis task was reformulated as a series of *one-vs-rest* binary classification problems. For each diagnostic class $c \in \{CN : 1, MCI : 2, AD : 3\}$, a binary label vector was constructed as

$$y_i^{(c)} = \begin{cases} 1, & \text{if } y_i = c, \\ 0, & \text{otherwise.} \end{cases}$$

This formulation allows independent analysis of discriminative features for each diagnostic group while maintaining clinical interpretability.

B. Handling Class Imbalance

The ADNI dataset exhibits class imbalance, particularly for the AD group. Rather than resampling the data, a cost-sensitive learning strategy was adopted to preserve the biological plausibility of biomarkers. Class imbalance was addressed through a multi-level strategy. First, a one-vs-rest formulation was adopted to enable class-specific modelling. Stratified cross-validation preserved class proportions across folds. In addition, cost-sensitive learning was applied by weighting the loss function inversely proportional to class frequency. For Elastic Net LR and RF classifiers, class-dependent weights were computed using the inverse class frequency in the training set. Let N denote the total number of training samples and N_k the number of samples belonging to class $k \in \{0, 1\}$. The weight assigned to class k is defined as: $w_k = \frac{N}{2N_k}$.

These weights were incorporated into the loss function, yielding the following weighted logistic regression objective:

$$\mathcal{L}(\boldsymbol{\beta}) = -\sum_{i=1}^{N} w_{y_i} \left[y_i \log p_i + (1 - y_i) \log(1 - p_i) \right] + \lambda \mathcal{R}(\boldsymbol{\beta}),$$

where p_i denotes the predicted probability of the positive class and $\mathcal{R}(\boldsymbol{\beta})$ represents the Elastic Net regularisation term.

In Random Forest models, the same class weights were applied during node splitting, ensuring that impurity measures (e.g., Gini impurity) penalised misclassification of minority class samples more heavily.

For XGBoost, class imbalance was addressed using the `scale_pos_weight` parameter, which directly rescales the gradients of positive class samples during boosting. The parameter was computed as: $\texttt{scale\_pos\_weight} = \frac{N_{\text{neg}}}{N_{\text{pos}}}$, where N_{pos} and N_{neg} denote the number of positive and negative samples in the training data, respectively.

This weighting modifies the gradient of the logistic loss, effectively increasing the contribution of minority class samples during tree construction and optimisation.

C. Elastic Net Logistic Regression

Elastic Net regularised Logistic Regression (ENLR) was used to obtain sparse, stable feature importance estimates with correlated predictors, including neuroimaging measures, cognitive scores, and biochemical variables [5]. In such settings, standard ℓ_2-regularised Logistic Regression spreads weights across correlated features, whereas pure ℓ_1 regularisation can produce unstable feature selection. ENLR addresses these issues by combining both penalties.

For features $\mathbf{x}_i \in \mathbb{R}^p$ and binary labels $y_i \in \{0, 1\}$, the class probability is: $P(y_i = 1 \mid \mathbf{x}_i) = \frac{1}{1+\exp(-\mathbf{w}^\top \mathbf{x}_i - b)}$, and the optimisation objective combines ℓ_1 and ℓ_2 penalties as:

$$\begin{aligned} \mathcal{L}(\mathbf{w}, b) = &-\sum_{i=1}^{N} \left[y_i \log p_i + (1 - y_i) \log(1 - p_i) \right] \\ &+ \lambda \left(\alpha \|\mathbf{w}\|_1 + \frac{1 - \alpha}{2} \|\mathbf{w}\|_2^2 \right), \end{aligned} \tag{1}$$

where $p_i = P(y_i = 1 \mid \mathbf{x}_i)$, $\lambda > 0$ controls regularisation power, and $\alpha \in [0, 1]$. A `saga` solver with $\alpha = 0.5$ was used

for optimisation. Finally, Importance$_j = |w_j|$, was used for feature importance and averaged across cross-validation folds.

D. Random Forest

Random Forest (RF) was employed to capture nonlinear relationships and feature interactions. An ensemble of $M = 300$ trees $\{T_1, \ldots, T_M\}$ was trained using bootstrap aggregation and random feature selection, where binary predictions were obtained by majority voting: $\hat{y} = \arg\max_{c\in\{0,1\}} \sum_{m=1}^{M} \mathbb{I}\left(T_m(\mathbf{x}) = c\right)$, with $\mathbb{I}(\cdot)$ denoting the indicator function.

For a feature f_j, feature importance was quantified using Mean Decrease in Impurity (MDI):

$$\text{Imp}(f_j) = \frac{1}{M} \sum_{m=1}^{M} \sum_{t\in T_m} \mathbb{I}(v_t = f_j) \cdot \Delta I_t, \tag{2}$$

where v_t denotes the feature selected at node t, and ΔI_t represents the reduction in impurity achieved by the split. Finally, the impurity was measured using Gini impurity: $I_{\text{Gini}} = 1 - \sum_{c=1}^{C} p_c^2$. The proportion of samples belonging to class c at the node is denoted by p_c.

E. Extreme Gradient Boosting

Unlike RF, which trains trees independently, XGBoost builds trees sequentially, with each new tree correcting the errors of the existing ensemble [31]. Given a training dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, XGBoost models the prediction as an additive function: $\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i)$ with each $f_k \in \mathcal{F}$ representing a regression tree and K is the total number of trees.

The training objective minimised by XGBoost was defined as: $\mathcal{L} = \sum_{i=1}^{N} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$, where $\ell(\cdot)$ and $\Omega(\cdot)$ denoted loss function and regularization to control model complexity. These were calculated as: $\ell(y_i, \hat{y}_i) = -y_i \log p_i - (1 - y_i) \log(1 - p_i)$, and $\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$,, where $p_i = \sigma(\hat{y}_i)$ with $\sigma(\cdot)$ is the sigmoid function, T is the number of leaf nodes and w_j is the weight associated with leaf j.

Finally, feature importance was computed by aggregating the gain achieved by splits involving each feature across all trees. For a feature f_j, the importance score is given by: $\text{Imp}(f_j) = \sum_{t\in\mathcal{T}} \mathbb{I}\big(v_t = f_j\big) \cdot \text{Gain}_t$, where $\mathcal{T}$ is the set of all split nodes, v_t is the feature used at node t, and Gain_t is the reduction in loss resulting from the split. Importance vectors were averaged across cross-validation folds.

IV. Explainability and Interpretability

Explainability and interpretability are essential for clinical decision-support systems, especially in the analysis of neurodegenerative diseases. This study addresses interpretability at both the data and model levels using feature correlation analysis and SHAP-based global explainability to reveal feature relationships, model behaviour, and biomarker relevance.

A. Feature Correlation Analysis

To assess dependencies and multicollinearity among ADNI features, pairwise Pearson correlations were computed on the preprocessed numeric feature set as: $\rho_{ij} = \frac{\text{cov}(f_i, f_j)}{\sigma_{f_i}\sigma_{f_j}}$, where $\text{cov}(\cdot)$ denotes covariance and σ_f represents the standard deviation of feature f.

The resulting correlation matrix shows linear relationships and redundancy among cognitive, imaging, and biological biomarkers in multimodal clinical data. ENR effectively manages correlated predictors through grouped feature selection using ℓ_1 and ℓ_2 regularisations [5], while tree-based ensembles such as RF and XGBoost remain robust to multicollinearity via split-based learning and ensemble averaging [30].

B. SHAP-Based Global Explainability

Using the XGBoost classifier, SHAP [40] provides model-level explainability to interpret model predictions. It is grounded in cooperative game theory and attributes each feature's contribution to a model's output ($\hat{y}(\mathbf{x})$) in a manner that satisfies local accuracy, consistency, and missingness properties. The decomposition of model output is done as: $\hat{y}(\mathbf{x}) = \phi_0 + \sum_{j=1}^{d} \phi_j$, where ϕ_0 is the expected model output over the background distribution and ϕ_j is the contribution of feature f_j. And the global feature importance was computed as the mean absolute SHAP value of all samples: $\text{GlobalImp}(f_j) = \frac{1}{N} \sum_{i=1}^{N} \left|\phi_j^{(i)}\right|$. This importance highlights the most influential biomarkers driving model predictions across diagnostic groups.

V. Results and Discussions

We analyse model behaviour and feature relevance using both global and local interpretability methods, including Top-K feature importance, cross-model agreement, feature correlations, and SHAP summary plots. These analyses identify key predictors, evaluate consistency across models, and explain feature contributions at both overall and individual levels, with feature definitions provided in Table I.

A. Top-K Feature Importance

Top-K feature importance was analysed across CN, MCI, and AD using Elastic Net LR, RF, and XGBoost (Fig. 1). For CN, CDGLOBAL consistently dominates, supported by memory (CDMEMORY, DXMPTR5, DXMPTR2) and functional measures (FAQFINAN, MMSCORE), reflecting preserved cognition with subtle deficits. In MCI, models agree on CDGLOBAL, FAQTOTAL, CDMEMORY, and CDRSB, indicating a transition marked by combined cognitive and early functional decline, reinforced by global severity scores (TOTAL13, TOTSCORE). For AD, importance shifts toward clinical severity and functional impairment, with CDRSB and TOTAL13 leading, and additional contributions from communication, judgment, and informant-based measures. Overall, disease progression shows a clear transition from global cognitive features to functional and severity-driven predictors, with a stable core of key features across models.

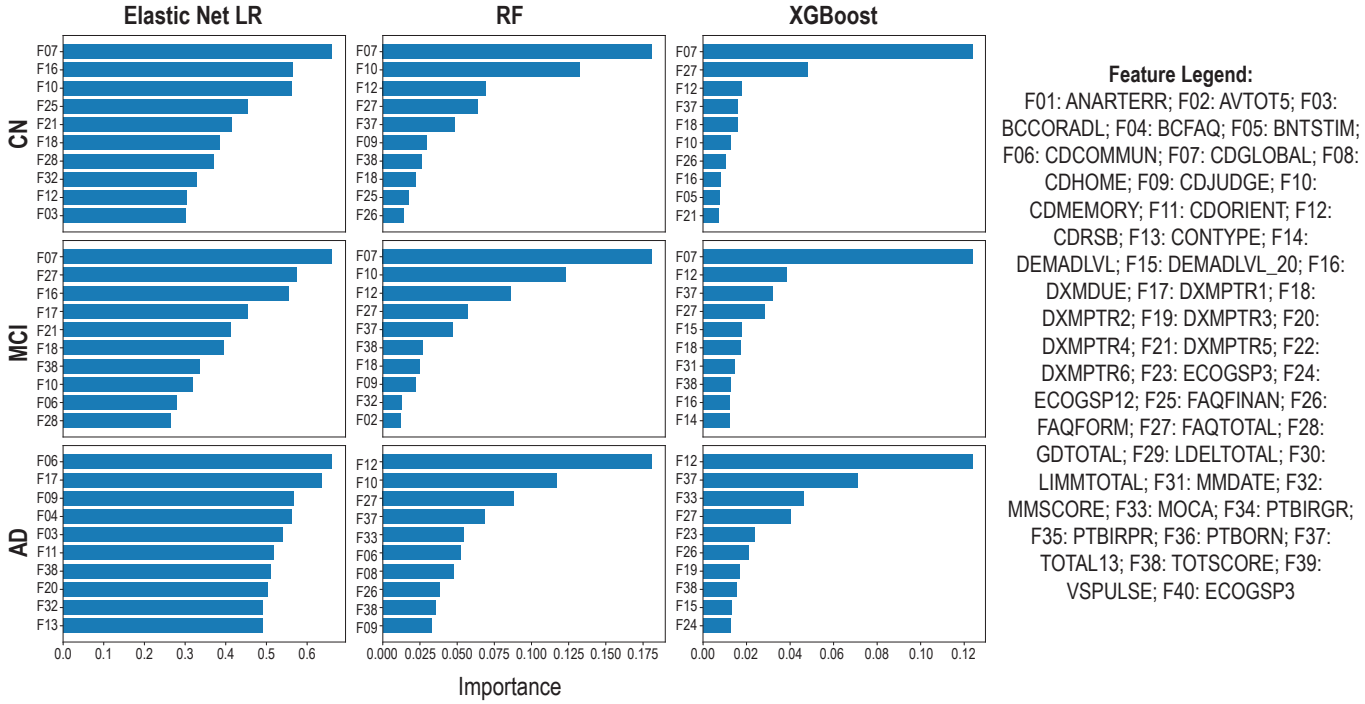


Fig. 1: Top-10 feature importance across diagnostic stages and models. Rows correspond to CN: Cognitively Normal, MCI: Mild Cognitive Impairment, AD: Alzheimer's Disease; columns correspond to LR: Elastic Net Logistic Regression, RF: Random Forest, and XGBoost: Extreme Gradient Boosting.

B. Cross-Model Feature Agreement Heatmap for AD

Figure 2 presents a heatmap of the Top 10 AD features, where brighter colours indicate higher importance. It reveals a clear distinction between LR Elastic Net and tree-based models (RF, XGBoost), with LR assigning consistently higher importance values. Despite this difference, all models converge on a stable set of AD-related features centred on clinical severity and functional impairment, particularly CDRSB and domain-level CDR measures. Cognitive indicators (CDORIENT, MMSCORE) and complaint-based features (DXMPTR1, DXMPTR4) provide additional support. Overall, the consistency across models suggests these features represent robust AD signals rather than model-specific artefacts.

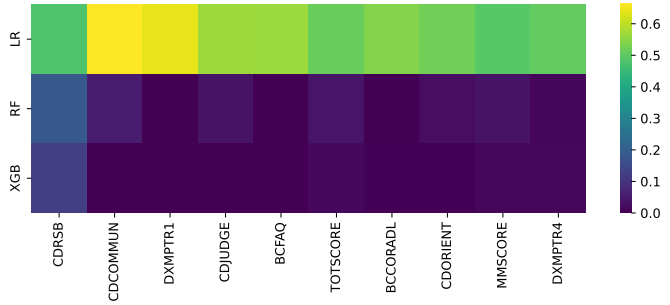


Fig. 2: Cross-model feature agreement for the AD class (LR Elastic Net, RF, XGBoost).

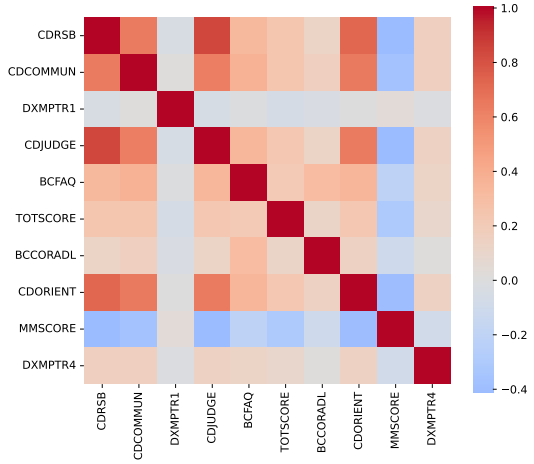
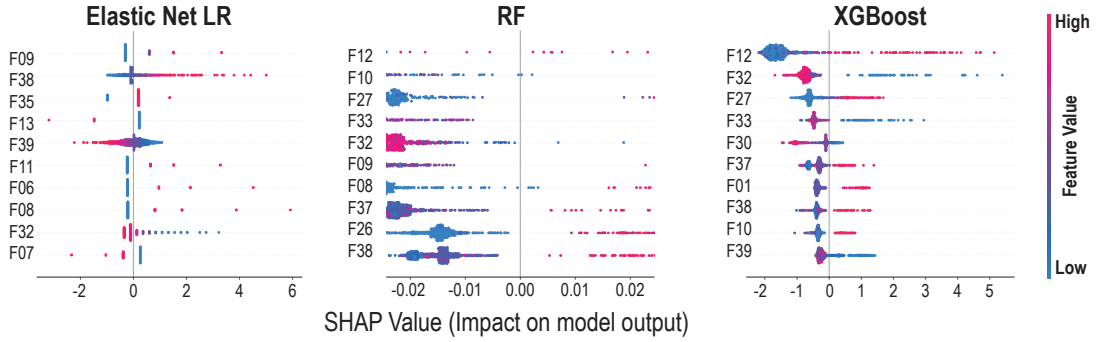


Fig. 3: Correlation matrix of cross-model consensus features for the AD class.

C. Correlation Matrix of Cross-Model Consensus Features

Figure 3 shows that agreed AD features form a strong severity/function cluster, with CDRSB, CDCOMMUN, CDJUDGE, and CDORIENT highly correlated, capturing global clinical impairment. TOTSCORE and BCFAQ moderately align with this cluster, whereas MMSCORE is negatively correlated, reflecting declining cognitive performance as severity increases. In contrast, DXMPTR1 shows weak correlations, indicating



Feature Legend: F01: ANARTERR; F02: AVTOT5; F03: BCCORADL; F04: BCFAQ; F05: BNTSTIM; F06: CDCOMMUN; F07: CDGLOBAL; F08: CDHOME; F09: CDJUDGE; F10: CDMEMORY; F11: CDORIENT; F12: CDRSB; F13: CONTYPE; F14: DEMADLVL; F15: DEMADLVL_20; F16: DXMDUE; F17: DXMPTR1; F18: DXMPTR2; F19: DXMPTR3; F20: DXMPTR4; F21: DXMPTR5; F22: DXMPTR6; F23: ECOGSP3; F24: ECOGSP12; F25: FAQFINAN; F26: FAQFORM; F27: FAQTOTAL; F28: GDTOTAL; F29: LDELTOTAL; F30: LIMMTOTAL; F31: MMDATE; F32: MMSCORE; F33: MOCA; F34: PTBIRGR; F35: PTBIRPR; F36: PTBORN; F37: TOTAL13; F38: TOTSCORE; F39: VSPULSE; F40: ECOGSP3

Fig. 4: SHAP summary plot showing the distribution of SHAP values for the Top-10 most influential features by mean |SHAP| for LR Elastic Net (left), RF (middle) and XGBoost (Right).

TABLE I: Definitions of clinical, cognitive, and functional features used in the analysis.

Feature	Description
CDGLOBAL	Global Clinical Dementia Rating
CDMEMORY	Clinical Dementia Rating (CDR) Memory Score
CDRSB	Clinical Dementia Rating Sum of Boxes (CDR-SB)
CDJUDGE	Clinical Dementia Rating Judgment and Problem Solving Score
CDORIENT	Clinical Dementia Rating Orientation Score
CDCOMMUN	Clinical Dementia Rating Community Affairs Score
CDHOME	Clinical Dementia Rating Home and Hobbies Score
MMSCORE	Mini-Mental State Examination (MMSE) Total Score
MMDATE	Mini-Mental State Examination Date Score
MOCA	Montreal Cognitive Assessment Score
TOTAL13	ADAS-Cognitive Total Score (ADAS-13)
TOTSCORE	ADAS-Cognitive Total Score (ADAS-11)
LDELTOTAL	Logical Memory Delayed Recall (Total Units Recalled)
LIMMTOTAL	Logical Memory Immediate Recall (Total Units Recalled)
AVTOT5	Neuropsychological Battery Trial 5 Total
ANARTERR	Neuropsychological Battery ANART Total Errors
BNTSTIM	Neuropsychological Battery Total Semantic Cues Given
FAQTOTAL	Functional Assessment Questionnaire Total Score
FAQFINAN	FAQ Managing finances (checks, bills, balancing)
FAQFORM	FAQ Assembling tax records or business papers
BCFAQ	Clinically relevant worsening in activities of daily living (FAQ)
BCCORADL	Informant-corroborated change in activities of daily living
DXMDUE	Diagnostic summary indicating MCI
DXMPTR1	Subjective memory complaint
DXMPTR2	Informant-reported memory complaint
DXMPTR3	Normal general cognitive function
DXMPTR4	Normal activities of daily living
DXMPTR5	Objective memory impairment for age and education
DXMPTR6	Not demented by diagnostic criteria
GDTOTAL	Geriatric Depression Scale Total Score
DEMADLVL	Pre-Amyloid PET perceived dementia risk level
DEMADLVL_20	Post-Amyloid PET perceived dementia risk level
CONTYPE	Consent tracking type
PTBORN	Parent born in the United States or Canada
PTBIRPR	Parent birth region
PTBIRGR	Parent birth group
VSPULSE	Seated pulse rate (beats per minute)

independent contribution, and BCCORADL and DXMPTR4 have mild associations. Overall, AD feature agreement is driven by a correlated severity-based group, explaining consistent yet slightly varied feature prioritisation across models.

D. SHAP Summary Plots

Figure 4 shows SHAP summary plots for Elastic Net LR, RF, and XGBoost, capturing feature impact, direction, and variability across samples. Elastic Net LR shows clear linear, monotonic effects, with features such as TOTSCORE, MMSCORE, CDGLOBAL, and CDORIENT exhibiting consistent directional influence. RF highlights CDRSB as a stable, positive contributor, while MMSCORE and MOCA have negative effects, and functional features exhibit moderate variability with more diffuse importance. XGBoost shows the widest SHAP ranges and strongest non-linear behaviour, with CDRSB as the dominant positive driver and MMSCORE inversely related to predictions, alongside heterogeneous effects from other features. Overall, LR reflects structured linear relationships, RF provides stable but distributed contributions, and XGBoost captures complex, non-linear interactions.

VI. Conclusion

This study investigated AD staging using ADNI-derived clinical, cognitive, and functional variables with Elastic Net LR, RF, and XGBoost, emphasising interpretable, robust feature relevance over accuracy alone. Across CN, MCI, and AD tasks, the models identified a clinically consistent progression pattern: CN is primarily characterised by preserved global cognition with subtle memory and function-related signals; MCI reflects a transitional profile combining cognitive impairment and emerging functional decline; and AD is dominated by clinical severity and loss of daily functioning. Importantly, cross-model feature agreement revealed a stable explanatory core for AD that includes CDRSB and multiple CDR-domain and functional measures, suggesting these predictors are model-stable rather than artefacts of a single classifier. Correlation analysis further showed that many consensus features form a tightly correlated cluster of severity and function. SHAP-based summaries complemented these findings by confirming

consistent directional relationships among major predictors and highlighting that boosted trees concentrate attribution on fewer features than linear models. Taken together, the results support the central premise that reliable interpretability in ADNI tabular modelling requires pairing feature importance with cross-model robustness and correlation structure.

REFERENCES

[1] G. Livingston, J. Huntley, A. Sommerlad *et al.*, "Dementia prevention, intervention, and care: 2020 report of the lancet commission," *The Lancet*, vol. 396, no. 10248, pp. 413–446, 2020.

[2] M. Bracher-Smith, F. Melograna, B. Ulm, C. Bellenguez, B. Grenier-Boley, D. Duroux, A. J. Nevado, P. Holmans, B. M. Tijms, M. Hulsman *et al.*, "Machine learning in alzheimer's disease genetics," *Nature communications*, vol. 16, no. 1, p. 6726, 2025.

[3] S. Kumar, I. Oh, S. Schindler, A. M. Lai, P. R. Payne, and A. Gupta, "Machine learning for modeling the progression of alzheimer disease dementia using clinical data: a systematic literature review," *JAMIA open*, vol. 4, no. 3, p. ooab052, 2021.

[4] F. M. D'Amore *et al.*, "Explainable machine learning on clinical features to predict and differentiate alzheimer's progression by sex: Toward a clinician-tailored web interface," *Journal of the Neurological Sciences*, vol. 468, p. 123361, 2025.

[5] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 2, pp. 301–320, 2005.

[6] C. Molnar, *Interpretable Machine Learning*. Leanpub, 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[7] V. Viswan, N. Shaffi, M. Mahmud *et al.*, "Explainable artificial intelligence in alzheimer's disease classification: A systematic review," *Cogn. Comput.*, vol. 16, no. 1, pp. 1–44, 2024.

[8] I. Oiza-Zapata and A. Gallardo-Antolín, "Alzheimer's disease detection from speech using shapley additive explanations for feature selection and enhanced interpretability," *Electronics*, vol. 14, no. 11, 2025.

[9] C.-H. Huang, F. A. Batarseh, and A. Ullah, "Evaluating the effectiveness of machine learning for alzheimer's disease prediction using applied explainability," *Biophysica*, vol. 5, no. 4, 2025.

[10] M. B. T. Noor, N. Z. Zenia, M. S. Kaiser, S. A. Mamun, and M. Mahmud, "Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of alzheimer's disease, parkinson's disease and schizophrenia," *Brain Inform.*, vol. 7, no. 1, p. 11, 2020.

[11] J. Ruiz, M. Mahmud *et al.*, "3d densenet ensemble in 4-way classification of alzheimer's disease," in *Proc. BI*, 2020, pp. 85–96.

[12] N. Shaffi, F. Hajamohideen *et al.*, "Ensemble classifiers for a 4-way classification of alzheimer's disease," in *Proc. AII*, 2022, pp. 219–230.

[13] F. Hajamohideen, N. Shaffi, M. Mahmud *et al.*, "Four-way classification of alzheimer's disease using deep siamese convolutional neural network with triplet-loss function," *Brain Inform.*, vol. 10, no. 1, pp. 1–13, 2023.

[14] N. Shaffi *et al.*, "Performance evaluation of deep, shallow and ensemble machine learning methods for the automated classification of alzheimer's disease," *Int. J. Neural Syst.*, vol. 34, no. 07, p. 2450029, 2024.

[15] Y. Miah, C. N. E. Prima, S. J. Seema, M. Mahmud, and M. Shamim Kaiser, "Performance Comparison of Machine Learning Techniques in Identifying Dementia from Open Access Clinical Datasets," *Adv. Intell. Syst. Comput.*, vol. 1188, pp. 79 – 89, 2021.

[16] T. Chen, P. Su, Y. Shen, L. Chen, M. Mahmud, Y. Zhao, and G. Antoniou, "A dominant set-informed interpretable fuzzy system for automated diagnosis of dementia," *Front. Neurosci.*, vol. 16, 2022.

[17] M. Fabietti, M. Mahmud, A. Lotfi *et al.*, "Early detection of alzheimer's disease from cortical and hippocampal local field potentials using an ensembled machine learning model," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 2839–2848, 2023.

[18] S. Jesmin, M. S. Kaiser, and M. Mahmud, "Artificial and internet of healthcare things based alzheimer care during covid 19," in *Proc. BI*, 2020, pp. 263–274.

[19] F. Farhin, I. Sultana, N. Islam, M. S. Kaiser, M. S. Rahman, and M. Mahmud, "Attack Detection in Internet of Things using Software Defined Network and Fuzzy Neural Network," in *Proc. ICIEV icIVPR 2020*, 2020.

[20] F. Farhin *et al.*, "Secured smart healthcare system: Blockchain and bayesian inference based approach," *Adv. Intell. Syst. Comput.*, vol. 1309, pp. 455 – 465, 2021.

[21] F. Farhin, M. S. Kaiser, and M. Mahmud, "Towards Secured Service Provisioning for the Internet of Healthcare Things," in *Proc. AICT 2020*, 2020.

[22] S. Jahan, K. Abu Taher, M. S. Kaiser, M. Mahmud, M. S. Rahman, A. S. Hosen, and I.-H. Ra, "Explainable ai-based alzheimer's prediction and management using multimodal data," *Plos one*, vol. 18, no. 11, p. e0294253, 2023.

[23] V. Viswan, N. Shaffi, E. Malathy, G. Chemmalar Selvi, B. Kavitha, A. Abdesselam, S. Wang, P. N. Suganthan, I. A. Shezawi, and M. Mahmud, "Multimodal fusion and explainability of artificial intelligence models in alzheimer's disease detection," *Brain Informatics*, 2026.

[24] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of lime and shap in alzheimer's disease detection," *Brain Inform.*, vol. 11, no. 1, p. 10, 2024.

[25] J. Therriault, S. Janelidze, A. L. Benedet, N. J. Ashton, J. Arranz Martínez, A. Gonzalez-Escalante, B. Bellaver, D. Alcolea, A. Vrillon, H. Karim *et al.*, "Diagnosis of alzheimer's disease using plasma biomarkers adjusted to clinical probability," *Nature Aging*, vol. 4, no. 11, pp. 1529–1537, 2024.

[26] G. I. A. Calub, E. N. Elefante, J. C. A. Galisanao, S. L. B. G. Iguid, J. C. Salise, and S. V. Prado, "Eeg-based classification of stages of alzheimer's disease (ad) and mild cognitive impairment (mci)," in *Proc. BioSMART*, 2023, pp. 1–6.

[27] R. Xiao, X. Cui, H. Qiao, X. Zheng *et al.*, "Early diagnosis model of alzheimer's disease based on sparse logistic regression with the generalized elastic net," *Biomed. Signal Process. Control*, vol. 66, p. 102362, 2021.

[28] X. Cui, R. Xiao, X. Liu, H. Qiao, X. Zheng, Y. Zhang, and J. Du, "Adaptive lasso logistic regression based on particle swarm optimization for alzheimer's disease early diagnosis," *CChemom. Intell. Lab. Syst.*, vol. 215, p. 104316, 2021.

[29] R. Casanova, J. Harezlak *et al.*, "High dimensional classification of structural mri images via large scale regularization approaches," *Front. Neuroinform.*, vol. 5, p. 22, 2011.

[30] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[31] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. ACM SIGKDD*, 2016, pp. 785–794.

[32] B. Wang, R. Xie, W. Qi *et al.*, "Advancing alzheimer's disease risk prediction: development and validation of a machine learning-based preclinical screening model in a cross-sectional study," *BMJ open*, vol. 15, no. 2, p. e092293, 2025.

[33] C.-H. Huang, F. A. Batarseh, and A. Ullah, "Evaluating the effectiveness of machine learning for alzheimer's disease prediction using applied explainability," *Biophysica*, vol. 5, no. 4, p. 54, 2025.

[34] H. Alshamlan, A. Alwassel, A. Banafa, and L. Alsaleem, "Improving alzheimer's disease prediction with different machine learning approaches and feature selection techniques," *Diagnostics*, vol. 14, no. 19, p. 2237, 2024.

[35] F. Gu, S. Ma, X. Wang, J. Zhao, Y. Yu, and X. Song, "Evaluation of feature selection for alzheimer's disease diagnosis," *Front. Aging Neurosci.*, vol. 14, p. 924113, 2022.

[36] A. A. Soladoye, N. Aderinto, D. Osho, and D. B. Olawade, "Explainable machine learning models for early alzheimer's disease detection using multimodal clinical data," *Int. J. Med. Inform.*, p. 106093, 2025.

[37] G. P. McFall, L. Bohn, M. Gee, S. M. Drouin, H. Fah, W. Han, L. Li, R. Camicioli, and R. A. Dixon, "Identifying key multi-modal predictors of incipient dementia in parkinson's disease: a machine learning analysis and tree shap interpretation," *Front. Aging Neurosci.*, vol. 15, p. 1124232, 2023.

[38] T. Rai, Y. Shen, J. He *et al.*, "Understanding feature importance of prediction models based on lung cancer primary care data," in *Proc. IJCNN*, 2024, pp. 1–8.

[39] C. R. Jack Jr, M. A. Bernstein, N. C. Fox *et al.*, "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *J. Magn. Reson. Imaging*, vol. 27, no. 4, pp. 685–691, 2008.

[40] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.