STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 2026, pp 0–16.
Published online in International Academic Press (www.IAPress.org)

An Explainable Machine Learning–Based Web Tool for Alzheimer's Disease Diagnosis

Abdelaziz METTAHRI [1,*], Sokaina EL KHAMLICHI [1,2], M'barek EL HALOUI [1], Badia ETTAKI [1]

[1] *LyRICA: Laboratory of Research in Informatics, Data Sciences and Artificial Intelligence, School of Information Sciences, B.P. 6204, Rabat-Instituts, Rabat, Morocco*
[2] *Research Team in Science and Technology, Higher School of Technology of Laayoune, Ibn Zohr University, B.P. 3007, Laayoune, Morocco*

Abstract Alzheimer's disease (AD) is one of the most critical global health challenges, and early diagnosis remains essential for slowing disease progression and improving patient management. This study proposes an automated and interpretable machine learning framework for AD diagnosis that combines feature selection, explainable artificial intelligence (XAI), and deep learning approaches to support clinically transparent decision-making. A publicly available AD dataset containing 25 attributes was used. Feature importance was assessed using correlation analysis together with explainability-based importance measures derived from SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), leading to the identification of an optimal subset of features composed of Age, presence of the GeneticRiskFactorAPOE-ε4 allele and FamilyHistoryOfAlzheimer. Several classical machine learning models, including Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost) were evaluated alongside Deep Learning (DL) approaches based on Multilayer Perceptron (MLP) architectures. Two MLP-based models were investigated: a feature selection-based MLP (FS-MLP), trained on the selected features, and an encoder-based MLP, integrating deep learning-driven dimensionality reduction. SHAP and LIME were further employed to provide both global and local explanations of model predictions. The selected model was deployed in a web-based application enabling AD diagnosis and result interpretation. Performance evaluation using cross-validation, confidence intervals, and statistical significance testing showed that tree-based models achieved the best overall performance, with FS-XGBoost obtaining the highest Area Under the Curve (AUC) of 80.1%, a recall of 77.8% and a specificity of 68.8%, indicating a balanced sensitivity–specificity trade-off for screening applications. To enhance practical applicability, the selected model was integrated into a user-friendly web-based decision-support application for AD diagnosis and interpretation. The results demonstrate that combining feature selection, explainable AI, and machine learning techniques, can provide an effective and transparent framework for supporting early AD diagnosis in clinical settings.

Keywords Alzheimer's Disease Diagnosis, Machine Learning, Multilayer Perceptron, Feature Selection, Explainable Artificial Intelligence

DOI: 10.19139/soic-2310-5070-3624

1. Introduction

Alzheimer's Disease (AD), the primary cause of memory loss and changes in mental capacity and behaviour, is increasingly raising concerns at both the World Health Organization (WHO) and Alzheimer's Disease International (ADI). The WHO reported that more than 57 million cases of dementia were recorded in 2021. AD accounts for 60%-70% of dementia cases [1]. Despite AD is an incurable and irreversible disease, its progression can be reduced if it is identified early and treated with appropriate medicine and rehabilitation [2]. AD can be avoided as well by understanding risk factors and managing risk reduction [3].

*Correspondence to: Abdelaziz METTAHRI (Email: abdelaziz.mettahri@esi.ac.ma). LyRICA: Laboratory of Research in Informatics, Data Sciences and Artificial Intelligence, School of Information Sciences, B.P. 6204, Rabat-Instituts, Rabat, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

In recent years, Machine Learning (ML) and Deep Learning approaches have demonstrated remarkable performance across a wide range of medical diagnosis and prediction tasks [4, 5]. Motivated by these advances, numerous studies have explored the use of ML and DL techniques for the early detection and diagnosis of Alzheimer's disease (AD). In particular, the analysis of Magnetic Resonance Imaging (MRI) data using advanced DL architectures, such as Convolutional Neural Networks (CNNs), vision transformers and autoencoders, has yielded outstanding outcomes in AD classification and prediction [6–9]. This high performance is largely attributed to the richness of imaging data, which captures detailed structural brain changes associated with AD. Despite this advantage, such methods face practical limitations, including high acquisition cost, limited accessibility in primary care settings, and the need for specialized equipment and expertise, which restrict their large-scale deployment. Additionally, recent studies have investigated transformer-based models for analyzing speech and clinical text in dementia detection, enabling the extraction of linguistic and cognitive biomarkers associated with AD [10]. More recently, large language models (LLMs) have been employed to extract clinically relevant information from unstructured medical data, such as electronic health records and clinical notes, facilitating the identification of AD-related symptoms and risk factors [11]. In parallel, federated learning approaches have been proposed to enable collaborative model training across multiple clinical sites while preserving data privacy, which is particularly relevant for AD studies involving distributed patient data [12]. Other investigations used tabular datasets and traditional supervised ML such as Random Forest (RF), Light Gradient-Boosting Machine (LightGBM), eXtreme Gradient Boosting (XGBoost), Logistic Regression (LR), and Support Vector Machine (SVM) in AD prediction [13–16].

Despite these advances, the development of effective AD prediction models faces several challenges. Medical datasets often contain a large number of heterogeneous variables, making model training complex and limiting practical deployment in clinical settings. Moreover, many ML and DL models operate as black boxes, providing limited interpretability, which reduces their acceptance by healthcare professionals and hinders their integration into routine medical practice. Consequently, the absence of explainability represents a critical clinical barrier, as clinicians must be able to understand and trust model outputs before using them in practice. This limitation can be partly explained by the fact that prior work has predominantly focused on optimizing predictive performance, where feature selection is mainly used for dimensionality reduction rather than for providing interpretable insights. Moreover, the use of complex DL architectures inherently limits transparency and makes interpretability more difficult, leading explainability to be treated as an optional post-hoc analysis rather than a core design requirement. While most existing studies remain confined to research environments and focus primarily on predictive performance, often overlooking interpretability and real-world usability, this work adopts a comprehensive approach. It combines feature selection for dimensionality reduction, model optimization, and explainable artificial intelligence (XAI) techniques to develop an interpretable and clinically usable AD diagnosis tool.

The main objective of this study is to design and implement a high-performing and explainable ML-based model for automated AD prediction using a reduced set of relevant features. Feature selection is first applied to identify the most influential variables before model training. The resulting model is then integrated into a user-friendly web application aimed at assisting clinicians in AD diagnosis and decision interpretation. In contrast to prior studies that treat explainability as an optional extension, this work considers interpretability as a core design requirement, directly linked to clinical usability and deployment. Specifically, the integration of SHAP and LIME is not only intended to explain model predictions, but also to support clinical reasoning by highlighting the contribution of modifiable and non-modifiable risk factors. The contribution of the present paper lies in three main aspects. First, we provide a comprehensive comparison between classical machine learning (ML) algorithms and deep learning (DL) models for Alzheimer's disease (AD) diagnosis using tabular clinical data. Second, we integrate explainable artificial intelligence (XAI) techniques, namely SHAP and LIME, to enhance model interpretability, identify clinically relevant risk factors, and improve transparency in the decision-making process. Third, we translate the proposed framework into a practical web-based clinical decision-support system, thereby facilitating its use in real-world healthcare settings.

The rest of this paper is structured as follows: Section 2 describes the proposed methodology, Section 3 provides the outcomes of our study, Section 4 discusses our findings, Section 5 concludes this research and presents future works.

2. Methods

2.1. Dataset description

The dataset employed in our study is a public dataset extracted from Kaggle, composed of 74,283 patients [17]. The dataset used comprises 25 independent features providing insights about lifestyle, medical, demographic, and genetic factors. The dataset variable Alzheimer's Diagnosis with two output classes (Yes, No) indicates the target of our study. Table 1 provides additional details about the dataset.

Table 1. Dataset description

Attribute	Description	Data type
Age	The patient's age	Numerical
Gender	Male, Female	Categorical
Country	Patient country	Categorical
Employment Status	Employment Status (Employed, Retired, Unemployed)	Categorical
Social Engagement Level	Social Engagement Level (High, Medium, Low)	Categorical
Income Level	Income Level (High, Medium, Low)	Categorical
Marital Status	Marital Status (Married, Single, Widowed)	Categorical
Urban vs Rural Living	Urban, Rural	Categorical
Education Level	Education year, with a range of 0 to 19 years	Numerical
BMI	Body Mass Index of the patients	Numerical
Smoking Status	Smoking Status (Current, Former, Never)	Categorical
Alcohol Consumption	Alcohol consumption per week	Categorical
Diabetes	No, Yes	Categorical
Physical Activity Level	Level of weekly physical exercise	Categorical
Dietary Habits	Diet quality level (Healthy, Average, Unhealthy)	Categorical
Air Pollution Exposure	Air Pollution Exposure (High, Medium, Low)	Categorical
Stress Levels	Stress Levels (High, Medium, Low)	Categorical
Sleep Quality	Sleep quality level (Good, Average, Poor)	Categorical
Family History of Alzheimer's	Family history of Alzheimer's Disease, No, Yes	Categorical
Depression Level	Depression Level (High, Medium, Low)	Categorical
Cholesterol Level	No, Yes. with two levels High and Normal	Categorical
Hypertension	No, Yes	Categorical
Cognitive Test Score	Cognitive Test Score to assess level of cognition	Numerical
Genetic Risk Factor	Genetic Risk Factor (APOE-ε4 allele): No, Yes	Categorical
Alzheimer's Diagnosis	Diagnosis status for AD; Yes, No	Categorical

2.2. Data preprocessing and balancing

We have started with cleaning data, the first step of data processing, which revealed no missing values. Next, we applied feature encoding of categorical variables and scaling to standardize numerical features (using StandardScaler) to improve performance, stability, and training speed of the ML models. Following that, we reserved 80% of the AD dataset for model training, while the remaining 20% was used for testing.

The dataset exhibits class imbalance, with approximately 59% of samples belonging to the no AD class and 41% to the AD class. To address this imbalance without introducing data leakage, the Synthetic Minority Oversampling Technique (SMOTE) [18] was applied only to the training set after the train/test split to enhance class representation during training.

2.3. Feature selection and explainability techniques

To measure global feature importance, our study used three methods of feature selection known for their relevance, namely correlation analysis, SHAP values and an exploratory aggregation of LIME-based local explanations to provide an indicative global view of feature importance. It is important to note that LIME is inherently a local explanation method; therefore, the aggregation of LIME explanations is used in this study as an exploratory analysis rather than a rigorous or validated global feature importance technique. One of the most popular filter techniques is to compute the correlation matrix. A correlation matrix is a table displaying correlation coefficients which indicate the strength of relationships between variables. One of the most common correlation coefficients is the Pearson correlation coefficient which determines the strength of linear relationships between two variables. It is defined by the formula (1) [19].

$$R(i,j) = \frac{\operatorname{cov}(x_i, x_j)}{\sqrt{\operatorname{var}(x_i)\ \operatorname{var}(x_j)}} \tag{1}$$

Where,
x_i : is the i^{th} variable,
x_j : is the j^{th} variable,
$cov()$: represents the covariance
$var()$: represents the variance.

In addition to feature importance extraction, we introduce a quantitative evaluation of explainability consistency to assess the agreement and stability of interpretation methods. Specifically, the agreement between correlation analysis, SHAP and aggregation LIME explanations is measured using the Jaccard similarity coefficient, computed on the set of top-10 important features identified by each method. The Jaccard similarity is defined as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

Where A and B represent the sets of selected features from two explanation methods [20].

Furthermore, the stability of feature importance rankings across samples is evaluated using Spearman rank correlation, which quantifies the consistency of feature ordering between SHAP and LIME explanations. It is defined as:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{3}$$

Where d_i denotes the difference between the ranks of each feature in the two methods, and n is the total number of features [21]. These complementary metrics allow us to assess both overlap (Jaccard) and ranking consistency (Spearman), providing a more robust evaluation of explainability reliability. SHAP offers a comprehensible way to comprehend how dataset features affect models used to classify AD. It offers explanations that are both local and global. The SHAP value $\phi_j(x)$ for feature x_j measures the contribution of x_j to the discrepancy between the baseline prediction (the average prediction across all patients) and the model's prediction $f(x)$ [22].

LIME is a local explanation technique that is independent of models. It describes how each variable affects the outcome for a particular subject [22].

LIME is used in this study for two complementary purposes: local explanation and global feature importance. Locally, it explains individual predictions by approximating the black-box model with an interpretable surrogate model and assigning feature weights indicating their local contribution, either positive or negative, to each specific decision. Globally, LIME explanations are aggregated across multiple instances to estimate overall feature importance and derive a dataset-level feature ranking.

2.4. Classical ML and DL models

Our investigation leverages the strong capacity of both DL and classical ML algorithms for classification. In addition to DL models such as MLP and Autoencoder, we also employed several classical ML classifiers, including LR, K-Nearest Neighbors (KNN), Decision Tree (DT), RF, and XGBoost, to provide a comprehensive comparative evaluation.

2.4.1. *ML classifiers* :

Classical ML classifiers such as LR, KNN, DT, and RF remain widely used for tabular data due to their efficiency and interpretability [23]. Among these approaches, ensemble methods often achieve superior performance, with XGBoost representing an advanced gradient boosting technique that sequentially improves model predictions on structured datasets [25, 38].

2.4.2. *MLP* :

MLP is a particular form of the original Perceptron model. Its input and output layers are separated by one or more hidden layers; neurons are arranged in layers; connections are always made between lower and upper layers; and neurons within the same layer are not connected. The act of adjusting the connection weights to achieve a minimal difference between the network output and the intended output is known as learning for the MLP [26].

2.4.3. *Autoencoder* :

The autoencoder algorithm is a member of a unique family of artificial neural network-based dimensionality reduction techniques. It seeks to minimize the reconstruction error of an input in order to learn a compact representation [27]. Because of this, autoencoders are helpful for applications like noise elimination, feature extraction, and decreasing dimensionality.

In our study, we experimented and compared the performance obtained by two separate approaches in order to obtain the optimal model. On the one hand, the first approach consists of an encoder followed by a MLP. The encoder takes an input vector of 24 features and first applies a fully connected layer. A dropout layer is then applied to reduce overfitting. The encoder ends with a bottleneck dense layer of 5 neurons, which compresses the input into a low-dimensional latent representation. The latent representation produced by the encoder is then fed into an MLP classifier. Overall, the first approach combines dimensionality reduction through a bottleneck encoder with a deeper MLP for classification. On the other hand, the second approach follows a feature selection strategy, where the selected top optimal features are used as input. In this setting, multiple classifiers are employed, including an MLP as well as classical ML models such as LR, KNN, DT, RF, and XGBoost. The MLP takes an input vector of 5 five features selected in feature selection step. This model contains dense hidden layers. A dropout layer is applied after these layers to mitigate overfitting. The network ends with a single-neuron output layer, responsible for the final prediction. We employed the Keras Tuner library to perform hyperparameter optimization for the DL models, while RandomizedSearchCV was used to optimize the classical ML classifiers.

2.5. Evaluation metrics

To assess the performance of the outcomes, we employed the evaluation metrics obtained from the confusion matrix values, namely True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). These metrics consist of accuracy, precision, recall (sensitivity), specificity , F1 score, and AUC score [28, 29] (Table 2).

2.6. Proposed solution

Our main objective is to develop an explainable highly performing ML model-based web tool for clinical decision support. Accordingly, we developed two approaches based mainly on two different dimension reduction techniques and two MLP DL models, to determine the highly performing one, referred to by the optimal model as shown in Figure 1. In the first approach, a hybrid model (Encoder-MLP) created from an encoder and un MLP classifier. The encoder compresses and extracts the five most important variables from the entire features of the dataset received as input. Then, the MLP part makes a binary classification based on the encoder output. On the other hand, the

Table 2. Evaluation metrics

Metric	Formula	Description
Accuracy	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$	Accuracy of a classifier
Precision	$Precision = \frac{TP}{TP+FP}$	The model's ability to avoid false positives
Recall	$Recall = \frac{TP}{TP+FN}$	The model's ability to avoid false negatives
Specificity	$Specificity = \frac{TN}{TN+FP}$	The model's ability to correctly identify patients without the disease
F1-score	$F1_{Score} = 2 * \frac{Precision*Recall}{Precision+Recall}$	The harmonic mean of precision and recall
AUC	$AUC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$	The model's ability to distinguish between positive and negative classes

second approach primarily consists of a feature selection strategy, where the top optimal key features identified during the feature selection step are used as input for multiple classifiers, including MLP (FS-MLP), LR (FS-LR), KNN (FS- KNN), DT (FS-DT), RF (FS-RF), and XGBoost (FS- XGBoost). Besides, a web application with a user-friendly interface has been implemented based on the optimal model, to make the results of our work more practical and accessible to doctors and caregivers. This application provides four key functionalities: integration of an optimized pre-trained deep learning model, automated diagnosis based on this model, interpretable explanations of the diagnostic results, and a Mini-Mental State Examination (MMSE) [30] cognitive assessment to support and finalize the physician's diagnosis (Figure 1).

3. Results

This section details the results of our study.

3.1. Feature selection analysis

To determine the most pertinent attributes of AD dataset, we first carried out feature selection using a variety of techniques, namely correlation analysis, SHAP feature importance and aggregated LIME feature importance, as illustrated in Figure 2.

The feature selection step revealed three highly relevant features, namely Age, Genetic Risk Factor (APOE-ε4 allele) and Family History of Alzheimer's. To assess the consistency of the selected features across correlation analysis, SHAP, and aggregated LIME, we evaluated the agreement between methods using Jaccard similarity scores computed on the top-ranked features for each pair of methods. This allows identifying overlapping important variables. In addition, the stability of feature importance rankings was assessed using Spearman rank correlation, which measures the consistency of feature ordering across methods and samples. Furthermore, a complementary ablation study was conducted to evaluate the impact of using different numbers of selected features and to identify the optimal subset that provides a good balance between performance and interpretability.

3.1.1. Explainability evaluation :

Tables 3–5 present the top 10 most important features identified using three complementary approaches, namely SHAP, aggregated LIME, and correlation analysis. The comparison of these results enables the evaluation of consistency across methods and the identification of the most influential variables associated with AD prediction.

The results presented in Tables 3–5 show that key features such as Age, Genetic Risk Factor (APOE-ε4 allele), and Family History of Alzheimer's are consistently ranked among the most important variables across the three methods, highlighting their clinical relevance. However, a detailed assessment of the consistency between explanation methods is further provided in Table 6.

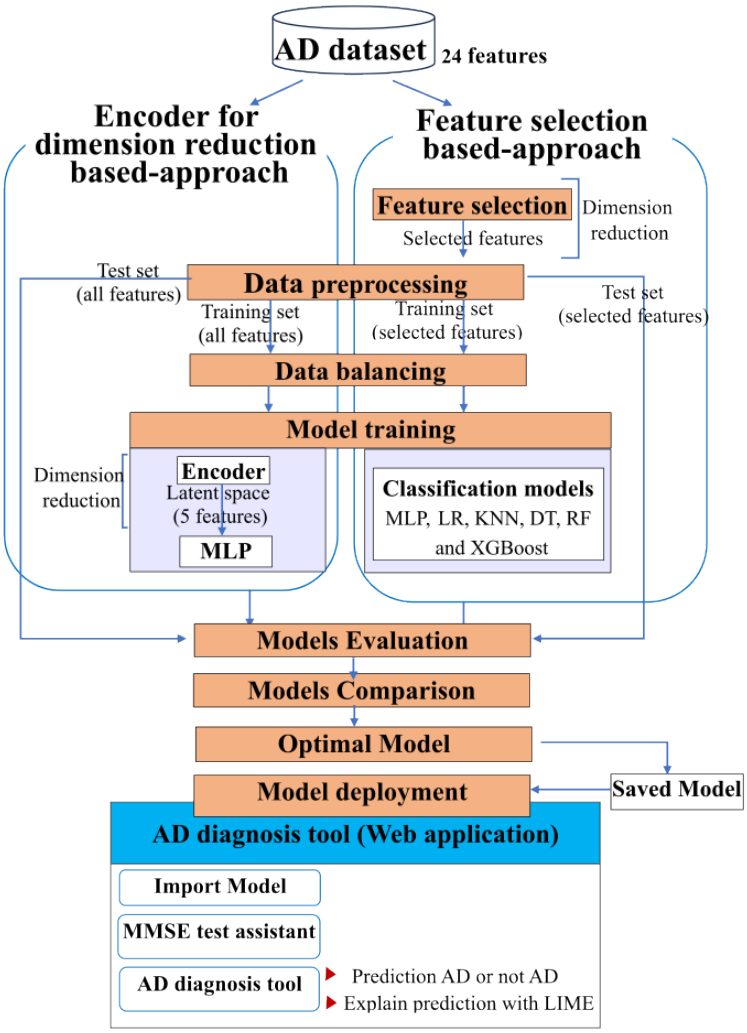

Figure 1. Overview of the proposed solution.

The Jaccard similarity scores indicate moderate-to-high overlap between the top-ranked features, with perfect agreement for the Top-3 features (Jaccard = 1 across all comparisons) and consistent overlap for larger subsets (Jaccard $\approx$ 0.429 for Top-5 and up to 0.667 for Top-10). These results confirm the robustness of the most important features across methods. Furthermore, the Spearman correlation values demonstrate strong agreement in feature ranking, particularly between SHAP and aggregated LIME (ρ = 0.840), while moderate correlations with correlation analysis ($\rho \approx$ 0.568–0.572) highlight complementary differences in how feature importance is captured.

3.1.2. Ablation study :

Table 7 presents the results of the ablation study conducted to evaluate the impact of different feature subsets on model performance. A Multilayer Perceptron (MLP) classifier with fixed architecture and hyperparameters was used as a baseline model to ensure a fair comparison across subsets. Several configurations were considered. The Minimal (Top-3) subset includes the three most important features (Age, GeneticRiskFactorAPOE-ε4 allele, and FamilyHistoryOfAlzheimer). The SHAP Top-5 and LIME Top-5 subsets correspond to the top five features identified by each respective explanation method. The Combined (Top-8) subset integrates features

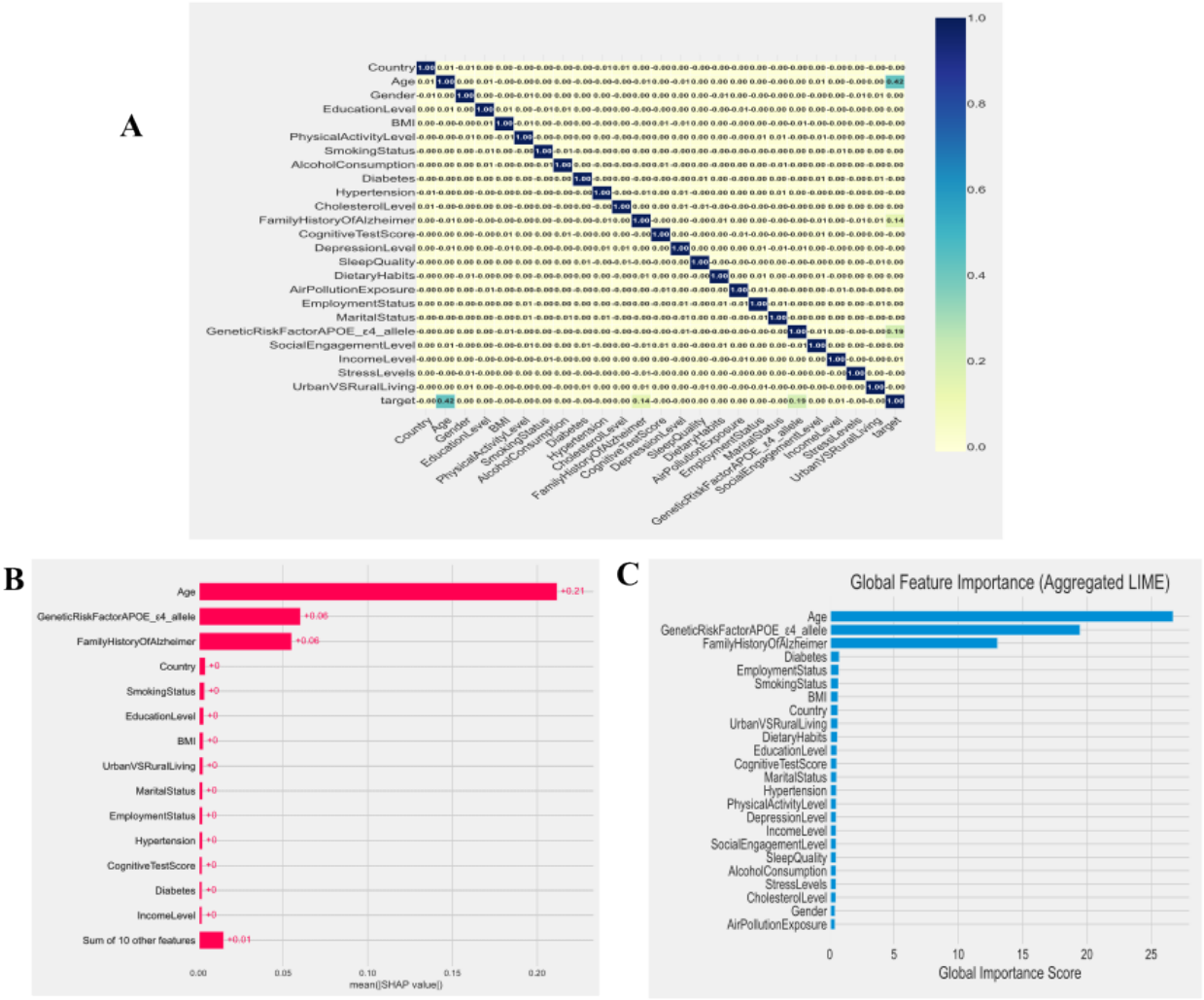

Figure 2. Features importance according to (A) correlation matrix (B) SHAP analysis and (C) aggregated LIME.

selected by SHAP, aggregated LIME, and correlation analysis, including Age, GeneticRiskFactorAPOE-ε4 allele, FamilyHistoryOfAlzheimer, Country, SmokingStatus, Diabetes, EmploymentStatus, and BMI. The Extended (Top-10) subset further expands this set by incorporating UrbanVSRuralLiving and EducationLevel. Finally, the full feature set includes all available variables. This setup enables the evaluation of the trade-off between feature dimensionality, predictive performance, and interpretability.

The results in Table 7 show that the Minimal (Top-3) subset achieves the best performance, with accuracy ($72.6 \pm 0.3\%$), precision ($72.0 \pm 0.3\%$), recall ($72.5 \pm 0.3\%$), F1-score ($72.1 \pm 0.3\%$), and AUC ($79.3 \pm 0.2\%$), indicating that a small set of key features captures most of the predictive signal. The SHAP Top-5 and LIME Top-5 subsets provide similar results, confirming the consistency of feature importance across methods. In contrast, the Combined (Top-8), Extended (Top-10), and full feature set show slightly lower performance, suggesting that adding more features introduces redundancy or noise.

3.1.3. Optimal feature set :

Table 3. Top 10 SHAP feature importances

Rank	Feature	Importance
1	Age	0.211742
2	GeneticRiskFactorAPOE_ϵ4_allele	0.060207
3	FamilyHistoryOfAlzheimer	0.055092
4	Country	0.004005
5	SmokingStatus	0.003568
6	EducationLevel	0.003063
7	BMI	0.002844
8	UrbanVSRuralLiving	0.002650
9	MaritalStatus	0.002378
10	EmploymentStatus	0.002317

Table 4. Top 10 aggregated LIME feature importances

Rank	Feature	Importance
1	Age	26.699874
2	GeneticRiskFactorAPOE_ϵ4_allele	19.455876
3	FamilyHistoryOfAlzheimer	13.043666
4	Diabetes	0.744831
5	EmploymentStatus	0.696529
6	SmokingStatus	0.673446
7	BMI	0.642367
8	Country	0.630330
9	UrbanVSRuralLiving	0.619529
10	DietaryHabits	0.584662

The identification of the optimal feature subset is supported by both explainability consistency and performance analysis. A perfect agreement across all methods is observed for the Top-3 features (Age, GeneticRiskFactorAPOE-ε4 allele, and FamilyHistoryOfAlzheimer) with Jaccard = 1.0, indicating that these variables are consistently recognized as the most important by SHAP, aggregated LIME, and correlation analysis. Beyond this subset, a decrease in agreement is observed, reflecting instability in the selection of lower-ranked features. Moreover, these top features exhibit the highest importance values across all methods, reinforcing their relevance. The ablation study further confirms this observation, showing that adding additional features does not improve model performance and may even lead to slight degradation. These findings demonstrate that the Top-3 feature subset provides an optimal balance between predictive performance, robustness, and interpretability.

3.2. Evaluation of test performance

The detailed performances of the evaluated models is presented in Table 8. Classical ML classifiers, including FS-LR, FS-KNN, FS-DT, FS-RF, and FS-XGBoost, were optimized using RandomizedSearchCV, while DL models (FS-MLP and Encoder-MLP) were optimized using Keras Tuner. This comparative evaluation allows assessing the effectiveness of both classical ML and DL approaches under optimized configurations.

Table 5. Top 10 correlation matrix feature importances

Rank	Feature	Importance
1	Age	0.419923
2	GeneticRiskFactorAPOE_ϵ4_allele	0.194484
3	FamilyHistoryOfAlzheimer	0.140885
4	IncomeLevel	0.006209
5	StressLevels	0.004671
6	UrbanVSRuralLiving	0.004104
7	EducationLevel	0.003732
8	Country	0.003694
9	SmokingStatus	0.003690
10	EmploymentStatus	0.003514

Table 6. Explainability consistency comparison

Comparison	Jaccard Top-3	Jaccard Top-5	Jaccard Top-10	Spearman Correlation (ρ)
SHAP vs Aggregated LIME	1	0.429	0.667	0.840
SHAP vs Correlation	1	0.429	0.667	0.568
Aggregated LIME vs Correlation	1	0.429	0.538	0.572

Table 7. Ablation study results

Features Subset	No. Features	Accuracy (mean $\pm$ std, %)	Precision (mean $\pm$ std, %)	Recall (mean $\pm$ std, %)	F1-score (mean $\pm$ std, %)	AUC (mean $\pm$ std, %)
Minimal (Top-3)	3	72.6 ± 0.3	72.0 ± 0.3	72.5 ± 0.3	72.1 ± 0.3	79.3 ± 0.2
SHAP Top-5	5	72.2 ± 0.2	71.5 ± 0.2	71.9 ± 0.2	71.6 ± 0.2	79.2 ± 0.4
LIME Top-5	5	72.3 ± 0.4	71.6 ± 0.4	72.0 ± 0.4	71.7 ± 0.4	79.2 ± 0.3
Combined (Top-8)	8	71.9 ± 0.4	71.2 ± 0.4	71.5 ± 0.5	71.3 ± 0.4	79.1 ± 0.3
Extended (Top-10)	10	71.9 ± 0.4	71.1 ± 0.4	71.4 ± 0.5	71.2 ± 0.4	78.9 ± 0.3
Full (all features)	24	71.3 ± 0.4	70.5 ± 0.5	70.5 ± 0.6	70.5 ± 0.5	78.3 ± 0.3

The results in Table 8 indicate that tree-based models, including DT, RF, and XGBoost, achieve the best overall performance across most evaluation metrics, with XGBoost and RF obtaining the highest AUC values of 80.1% and 80.0%, respectively, and strong recall ranging from 77.8% to 79.1%, which is critical in a clinical context to minimize missed AD cases. LR provides stable and well-balanced results across all metrics, although slightly below the top-performing models. In contrast, KNN achieves the highest specificity of 77.8% but lower recall, indicating a tendency to better identify negative cases at the expense of missed positives. The FS-MLP model performs competitively but does not outperform tree-based methods, while the Encoder-MLP shows the lowest performance, suggesting that autoencoder-based dimensionality reduction may lead to some loss of discriminative information. Overall, these findings highlight the effectiveness of tree-based models for tabular AD prediction.

This result is confirmed by Receiver Operating Characteristic (ROC) [31] curve analysis, as depicted in Figure 3.

3.3. Models performance comparison

3.3.1. Significance test of model performance differences :

Table 8. Performance of models

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)	AUC (%)	Best Hyperparameters
FS-LR	71.9	63.9	73.9	70.5	68.5	78.9	penalty elasticnet l1_ratio 0.11 C 0.0095
FS-KNN	71.0	66.1	61.4	77.8	63.7	78.3	weights distance n_neighbors 19 metric euclidean
FS-DT	72.5	63.6	78.4	68.4	70.2	79.9	max_depth 7 min_samples_split 15 min_samples_leaf 2
FS-RF	72.5	63.4	79.1	67.9	70.4	80.0	n_estimators 500 max_depth 5 min_samples_split 2 min_samples_leaf 2
FS-XGBoost	72.6	63.8	77.8	68.8	70.1	80.1	n_estimators 200 max_depth 3 learning_rate 0.1 subsample 0.9
FS-MLP	71.9	62.8	78.7	67.1	69.8	79.8	layers [192, 192, 64] dropout [0.3, 0.2, 0.2] lr 0.0003
Encoder-MLP	70.4	62.7	70.3	70.5	66.3	74.4	encoder $[64 \to 5 \to 64]$ encoder dropout 0.2 MLP layers [256] MLP dropout 0.3 lr 0.002

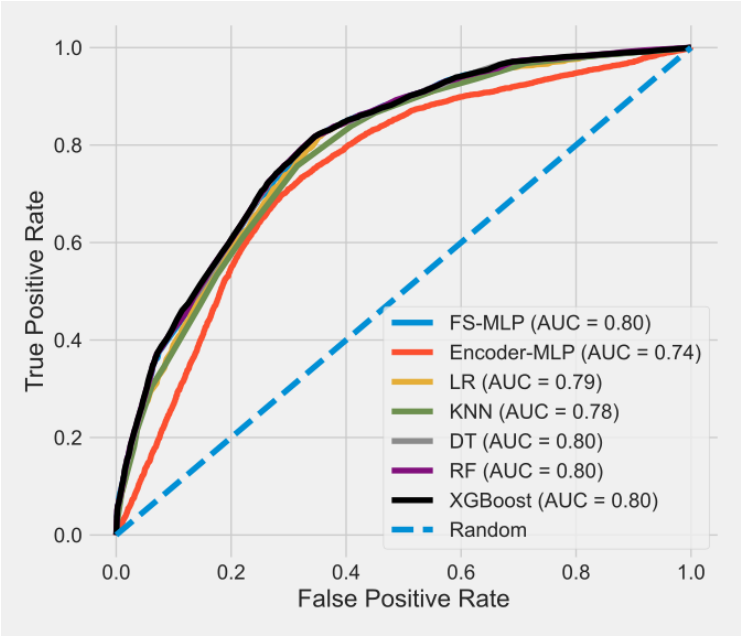

Figure 3. Performance comparison of classifiers through (A) confusion matrix (B) AUC metric and ROC curve.

A comprehensive comparison of model performance is provided based on 5-fold cross-validation results, including accuracy, precision, recall, specificity, F1-score, and AUC reported as mean ± standard deviation, as summarized in Table 9. In addition to these metrics, AUC 95% confidence intervals and p-values relative to the best-performing model are included to assess the statistical significance of performance differences. This analysis enables a rigorous evaluation of both predictive effectiveness and the reliability of observed differences across models.

Table 9. Performance Comparison and Significance Testing of Models

Model	**Accuracy (mean ± std, %)**	**Precision (mean ± std, %)**	**Recall (mean ± std, %)**	**Specificity (mean ± std, %)**	**F1 (mean ± std, %)**	**AUC (mean ± std, %)**	**AUC 95% CI (%)**	**p-value vs best AUC**
FS-XGBoost	72.1 ± 0.3	63.3 ± 0.5	77.2 ± 1.7	68.5 ± 1.3	69.6 ± 0.5	79.5 ± 0.3	[79.1, 79.9]	NaN
FS-DT	72.0 ± 0.4	63.2 ± 0.7	77.4 ± 1.3	68.2 ± 1.5	69.6 ± 0.3	79.5 ± 0.3	[79.1, 79.9]	0.725827
FS-RF	72.0 ± 0.3	63.2 ± 0.6	77.5 ± 0.9	68.2 ± 1.1	69.6 ± 0.2	79.5 ± 0.3	[79.1, 79.8]	0.295965
FS-MLP	71.8 ± 0.5	63.1 ± 1.2	77.0 ± 3.1	68.2 ± 2.8	69.3 ± 0.7	79.3 ± 0.3	[78.9, 79.7]	0.011649
FS-LR	71.3 ± 0.3	63.4 ± 0.4	72.1 ± 0.8	70.7 ± 0.7	67.5 ± 0.3	78.5 ± 0.4	[77.9, 79.0]	0.000180
FS-KNN	71.1 ± 0.4	65.9 ± 1.4	62.5 ± 4.3	77.1 ± 3.0	64.1 ± 1.6	77.9 ± 0.1	[77.7, 78.0]	0.000135
FS-Autoencoder	66.7 ± 2.0	57.1 ± 2.3	79.4 ± 4.1	57.7 ± 6.0	66.3 ± 0.9	73.4 ± 0.7	[72.6, 74.3]	0.000015

The results in Table 9 confirm that tree-based models, including XGBoost, DT and RF achieve the best overall performance, each obtaining an AUC of 79.5% with 95% confidence intervals ranging from 79.1% to 79.9% for XGBoost and DT, and from 79.1% to 79.8% for RF. The strong overlap of these intervals indicates that their predictive performance is statistically comparable, while their narrow width reflects stable behavior across folds. The associated p-values further support this observation, showing no statistically significant differences between these models. The FS-MLP model achieves a slightly lower AUC of 79.3% with a confidence interval from 78.9% to 79.7%, but the corresponding p-value indicates a statistically significant difference compared to the best-performing models. LR and KNN exhibit lower AUC values of 78.5% and 77.9% respectively, with confidence intervals with confidence intervals in lower ranges, reflecting reduced predictive performance. Additionally, KNN shows higher specificity of 77.1% but lower recall of 62.5%. The FS-Autoencoder model presents the lowest AUC of 73.4% with a wider confidence interval from 72.6% to 74.3%, indicating both weaker and less stable performance. Overall, these findings highlight the robustness and statistical reliability of tree-based models for AD prediction on tabular data.

3.4. Automated diagnosis tool

Our web application (Figure 4) provides a diagnosis functionality based on few variables, including age, genetic risk factor (APOE-ε4 allele) and family history of Alzheimer's disease, identified during the feature selection phase. The tool is designed as a clinical decision-support system rather than a standalone diagnostic solution, assisting healthcare professionals in the screening and early detection of AD. In this practical use case, healthcare professionals input these five attributes to generate an AD diagnosis for a patient. They can then activate the explanation module to interpret the model's decision, where a dedicated interface displays the contribution of each attribute to the predicted outcome. This interpretability enables informed clinical recommendations, particularly for modifiable risk factors such as hypertension. Additionally, the system integrates the MMSE as a cognitive assessment tool, assisting physicians in confirming and finalizing the diagnosis. Notably, the MMSE module operates independently from the predictive model and is intended as a complementary clinical reference rather than an input feature for the model classifier. More specifically, the MMSE score does not dynamically feed back into the AD classifier and does not influence the prediction output. This design reflects real clinical workflows, where cognitive test results and predictive model outputs are interpreted jointly by the physician rather than being automatically combined within a single model. Such separation ensures transparency, avoids unintended model bias, and allows clinicians to maintain full control over the final diagnostic decision.

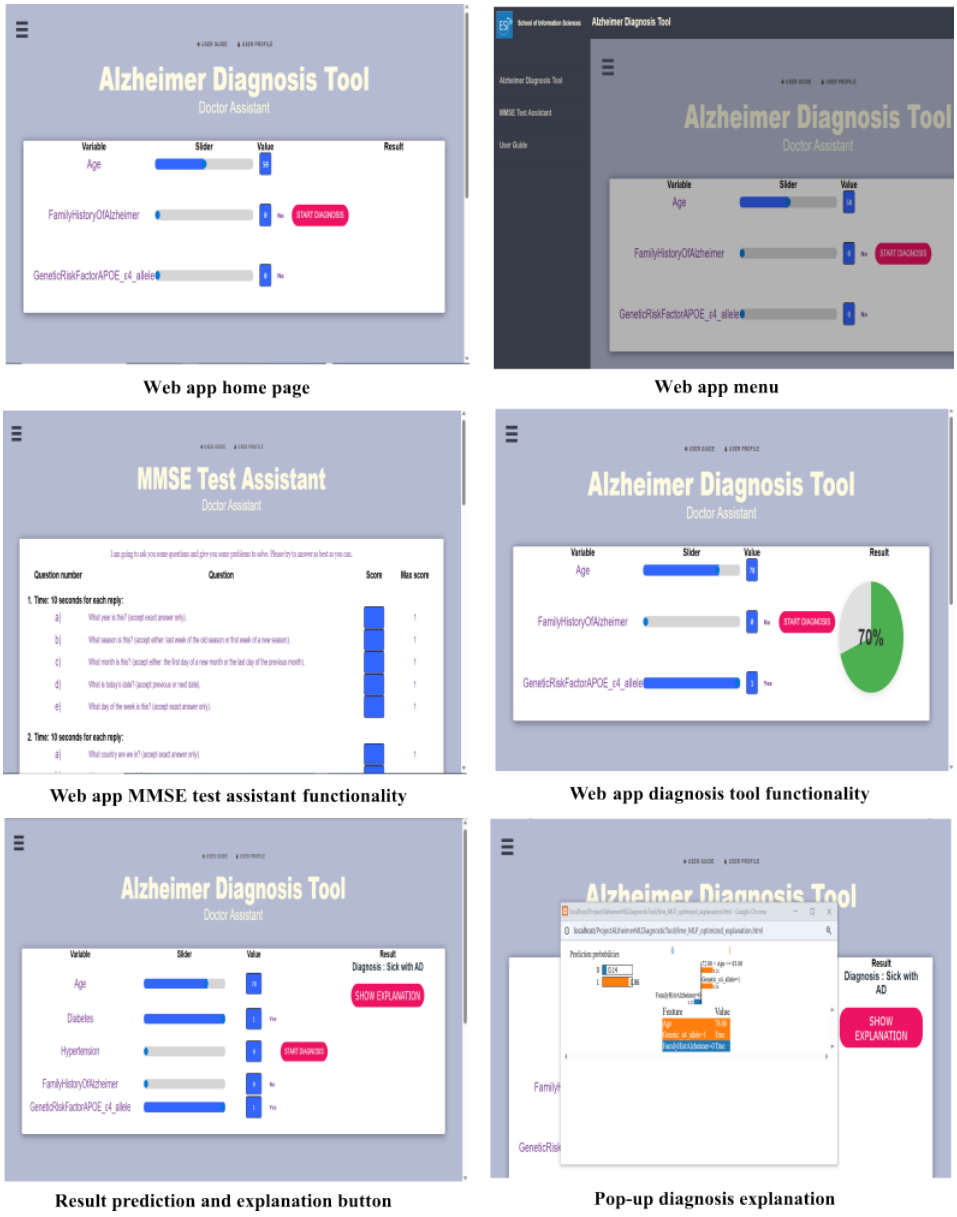

Figure 4. Screenshots of our AD diagnosis web application.

A demonstration video presenting the main features of the web application is freely accessible online via the URL [32].

Local model explanation of an instance can be offered by both SHAP and LIME (Figure 5). In our web application, we opted for LIME for its simple, clear, and detailed presentation.

4. Discussion

The outcomes of this study, especially the feature importance analysis results, highlight the important role that risk factors such as age, genetic risk factor (APOE-ε4 allele), family history of AD, hypertension, and diabetes play

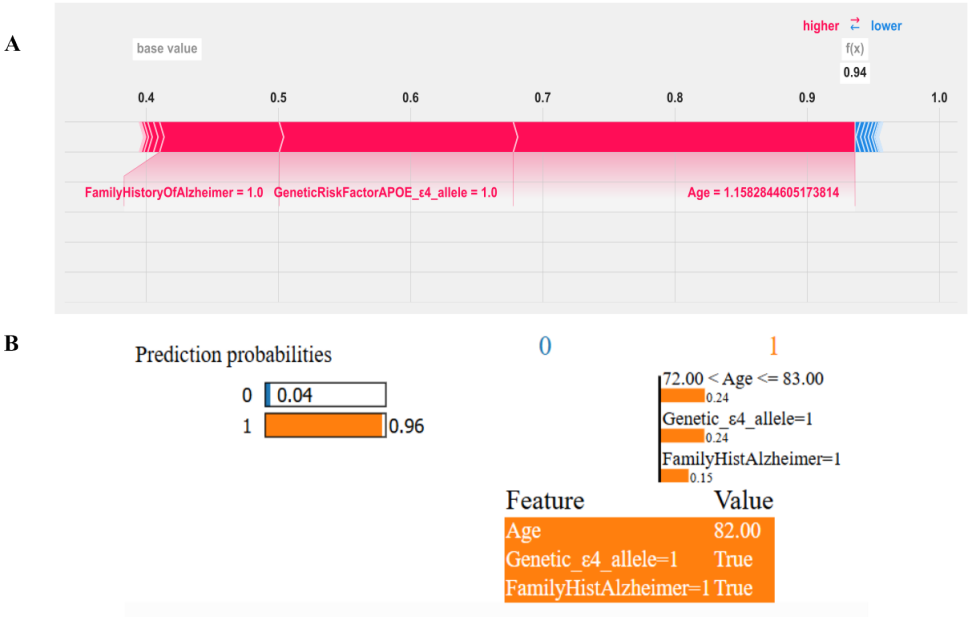

Figure 5. Local explanation according to (A) SHAP analysis and (B) LIME.

in the development of AD. This aligns with the study [33] which reported that according to meta-analyses, the frequency of AD increases threefold with a heterozygous ApoE $\varepsilon 4$ allele and fifteenfold with a homozygous ApoE $\varepsilon 4$ allele. Moreover, the study [34] conducted by research Group and Alzheimer's Disease Neuroimaging Initiative (ADNI) revealed that people who have a family history of AD (FHAD) are more likely to develop the condition themselves. An important contribution of this study is the quantitative evaluation of explainability robustness. A perfect agreement is observed for the Top-3 features (Jaccard = 1.0) across SHAP, aggregated LIME, and correlation analysis, while agreement decreases for larger feature subsets, reflecting variability in lower-ranked features. This result confirms that the most influential risk factors are consistently identified across different explanation methods. It should be noted that the aggregated LIME results are interpreted with caution, as they provide an approximate global perspective derived from local explanations and are not intended to replace more theoretically grounded methods such as SHAP. Ablation results show that the Minimal (Top-3) subset achieves the best performance, while adding more features does not improve results and may slightly degrade performance. This confirms that a compact set of key features captures the essential predictive signal. Furthermore, two of the top 5 significant factors found by our study, hypertension and diabetes, have a substantial link with the likelihood of acquiring AD, as stated in [35] and [36] , respectively.

This study completes and enhances our earlier work on ML-based AD diagnosis. Our prior research produced a performant and generalizable DT model through grid search optimization and validation using the 5-fold cross-validation approach [37]. The present study extends this contribution by exploring a new dataset and integrating classical ML models and DL approaches combined with explainability and statistical validation, including confidence intervals and significance testing. It also translates the final approach into an interactive web application for practical clinical support.

Performance comparison shows that tree-based models namely FS-DT, FS-RF, and FS-XGBoost achieve the best overall results, with statistically comparable performance supported by overlapping confidence intervals and non-significant p-values. Among them, FS-XGBoost achieves the highest AUC of 80.1% and maintains a balanced trade-off between recall of 77.8% and specificity of 68.8%, supporting its selection as the optimal model for deployment. This result highlights a controlled trade-off between sensitivity and specificity, where the model maintains high recall while preserving acceptable specificity, making it well suited for clinical screening scenarios.

From a clinical perspective, recall is especially important in AD prediction because it measures the model's ability to correctly identify patients with the disease and reduce false negatives. Avoiding missed diagnoses is crucial in medical decision making, as early detection can improve patient management and treatment outcomes [39]. However, the slightly lower specificity (68.8%) implies an increased number of false positives, which may lead to additional clinical examinations. This trade-off is acceptable in early screening contexts where missing true cases is more critical than over-referral. In practice, this trade-off can be controlled by adjusting the decision threshold of the classifier, allowing clinicians to balance sensitivity and specificity depending on the intended use case (screening vs confirmation). For instance, a lower decision threshold can be used in large-scale screening programs to maximize sensitivity, whereas a higher threshold may be adopted in clinical settings requiring higher specificity to reduce unnecessary follow-ups. A comparison with state-of-the-art studies (Table 10) shows that MRI and longitudinal MRI-based approaches generally achieve superior overall performance. In particular, study [9], which relies on genomic data, achieves competitive performance, highlighting the relevance of genetic information for AD prediction. However, such approaches typically require specialized data that are less accessible in routine clinical settings compared to the tabular features used in this study. Studies relying on the same AD tabular dataset report results comparable to ours. Notably, our approach achieved an improved recall while maintaining a competitive AUC score.

Table 10. Results comparison

Ref.	Dataset Type	Model	Accuracy %	Precision %	Recall %	Specificity %	F1-score %	AUC %
[6]	MRI data	DANMLP	93	-	94	92	-	95
[7]	MRI data	SVM	93	-	91	94	-	97
[8]	MRI data	CNN + DigitCapsule-Net	90	90	92	-	91	-
[9]	Genomic data	EDRNN	88	-	87	87	88	87
[13]	Longitudinal MRI data	RF	87	85	81	-	80	-
[15]	Longitudinal MRI data	RF	86	94	80	-	86	87
[14]	Tabular AD dataset [17]	Deep neural network	73	72	76	-	74	-
[16]	Tabular AD dataset [17]	Gradient boosting	73	72	73	-	72	-
		LR	71	71	70	-	70	-
		RF	72	72	68	-	68	-
Our study	Tabular AD dataset [17]	FS-XGBoost	72.6	63.8	77.8	68.8	70.1	80.1

To the best of our knowledge, relatively few studies have jointly incorporated feature selection and model explainability within a unified framework for AD diagnosis. Furthermore, only a limited number of existing works have extended their predictive models into deployable web-based applications. The comparison of our study with prior works highlights the novelty of our proposed framework, which integrates feature selection, data balancing, model explainability, and web application deployment.

Nevertheless, this study is not without limitations. Despite the application of data balancing techniques, the original class imbalance may still affect model generalizability. In addition, although the dataset contains a substantial number of samples for conventional machine learning methods, it remains relatively limited for deep learning approaches, which typically require much larger datasets to fully exploit their learning capabilities

5. Conclusion

This study investigated the contribution of feature selection and artificial intelligence techniques to the prediction of AD. The feature selection step identified the most significant risk factors associated with AD, improving model interpretability and predictive efficiency.Among the evaluated models, the FS-XGBoost achieved the best performance, with an accuracy of 72.6%, precision of 63.8%, recall of 77.8%, specificity of 68.8%, F1-score of 70%, and AUC of 80%. This optimal model was further integrated into a web-based tool, demonstrating its potential as a practical decision-support system for early AD diagnosis.Through our deployed AD diagnosis web application, we aimed to contribute toward bridging the gap between research and clinical practice by providing caregivers with an explainable AD prediction tool. However, further clinical validation remains necessary before considering

routine clinical adoption. Additionally, the explainability function facilitates the doctor's work by enabling clear interpretation of the model's decision and supporting informed clinical recommendations. Furthermore, the automated MMSE assistant helps efficiently validate and finalize the diagnosis. Both the predictive model and the MMSE module are intended to support rather than replace clinical judgment. For future work, we plan to address the issue of unbalanced datasets using several balancing data approaches and improving our study by conducting clinical validation with medical professionals. Also, our research can be expanded to make predictions using different types of data, such as genomic data and brain imaging. In addition, future work will investigate how XAI techniques can be systematically integrated into clinical decision-making protocols to support actionable recommendations.

REFERENCES

1. World Health Organization, *Dementia: Key facts*, [Internet], 2025. Available from: https://www.who.int/news-room/fact-sheets/detail/dementia
2. L. Clare and Jeon, *World Alzheimer Report 2025: Reimagining life with dementia – the power of rehabilitation*, [Internet], 2025. Available from: https://www.alzint.org/resource/world-alzheimer-report-2025
3. S. Long, C. Benoist, and W. Weidner, *World Alzheimer Report 2023: Reducing Dementia Risk: Never too early, never too late*, [Internet], 2023. Available from: https://www.alzint.org/resource/world-alzheimer-report-2023
4. Y. El Idrissi El-Bouzaidi, S. El Khamlichi, and O. Abdoun, *Diagnosis with Deep Learning and a Novel Pre-processing Strategy on a Large-Scale Dermoscopic Image Dataset*, Statistics, Optimization & Information Computing, vol. 15, no. 3, pp. 2069–2085, 2025. DOI: 10.19139/soic-2310-5070-3122.
5. L. Taidi and S. El Khamlichi, *A comprehensive review of artificial intelligence techniques for timely and accurate prediction of Down syndrome*, in *Agile Security in the Digital Era*, pp. 179–194, 2024.
6. Y. R. Qiang, S. W. Zhang, J. N. Li, Y. Li, and Q. Y. Zhou, *Diagnosis of Alzheimer's disease by joining dual attention CNN and MLP based on structural MRIs, clinical and genetic data*, Artificial Intelligence in Medicine, vol. 145, 102678, 2023. DOI: 10.1016/j.artmed.2023.102678.
7. P. M. Tuan, T. L. Phan, M. Adel, E. Guedj, and N. L. Trung, *AutoEncoder-based feature ranking for Alzheimer Disease classification using PET image*, Machine Learning with Applications, vol. 6, 100184, 2021. DOI: 10.1016/j.mlwa.2021.100184.
8. M. A. Bravo-Ortíz et al., *Transformers and capsule networks vs classical ML on clinical data for Alzheimer classification*, PeerJ Computer Science, vol. 11, e3208, 2025. DOI: 10.7717/peerj-cs.3208.
9. N. Mahendran and D. R. V. P. M, *A deep learning framework with an embedded-based feature selection approach for the early detection of Alzheimer's disease*, Computers in Biology and Medicine, vol. 141, 105056, 2022. DOI: 10.1016/j.compbiomed.2021.105056.
10. A. Balagopalan, B. Eyre, J. Robin, F. Rudzicz, and J. Novikova, *Comparing Pre-trained and Feature-Based Models for Prediction of Alzheimer's Disease Based on Speech*, Frontiers in Aging Neuroscience, vol. 13, 635945, 2021. DOI: 10.3389/fnagi.2021.635945.
11. I. Y. Oh et al., *Comparing Alzheimer disease phenotype extraction using rule-based natural language processing, GPT-4, Phi-4, LLaMA, and DeepSeek*, npj Dementia, vol. 1, no. 1, 37, 2025. DOI: 10.1038/s44400-025-00043-x.
12. K. Khalil et al., *Multidisciplinary Digital Publishing Institute*, Sensors, vol. 23, no. 19, 8272, 2023. DOI: 10.3390/s23198272.
13. C. Kavitha et al., *Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models*, Frontiers in Public Health, vol. 10, 853294, 2022. DOI: 10.3389/fpubh.2022.853294.
14. S. A. Mohamed et al., *Towards Transparent and Robust Early Diagnosis of Alzheimer's Disease: A Deep Learning and Visual Analytics Framework*, In Proceedings of Intelligent Methods, Systems, and Applications (IMSA), Giza, Egypt, pp. 340–345, 2025. DOI: 10.1109/IMSA65733.2025.11167773.
15. S. Demir and H. Selvitopi, *Early Diagnosis of Alzheimer's Disease Using Machine Learning Methods*, Procedia Computer Science, vol. 258, pp. 107–117, 2025. DOI: 10.1016/j.procs.2025.04.201.
16. B. Singh et al., *Evaluating Machine Learning Algorithms for Early Detection of Alzheimer's Disease*, In Proceedings of ICC-ROBINS, Coimbatore, India, pp. 510–515, 2025. DOI: 10.1109/ICC-ROBINS64345.2025.11086346.
17. A. Panday, *Alzheimer's Prediction Dataset (Global)*, [Internet], 2025. Available from: https://www.kaggle.com/datasets/ankushpanday1/alzheimers-prediction-dataset-global
18. N. V. Chawla et al., *SMOTE: Synthetic Minority Over-sampling Technique*, Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002. DOI: 10.1613/jair.953.
19. G. Chandrashekar and F. Sahin, *A survey on feature selection methods*, Computers & Electrical Engineering, vol. 40, no. 1, pp. 16–28, 2014. DOI: 10.1016/j.compeleceng.2013.11.024.
20. W. Maalej, M. Ellmann and R. Robbes, *Using contexts similarity to predict relationships between tasks*, Journal of Systems and Software, vol. 128, pp. 267–284, 2017. DOI: 10.1016/j.jss.2016.11.033.
21. K. Ali Abd Al-Hameed, *Spearman's correlation coefficient in statistical analysis*, International Journal of Nonlinear Analysis and Applications, vol. 13, no. 1, pp. 3249–3255, 2022. DOI: 10.22075/ijnaa.2022.6079.
22. A. M. Salih et al., *A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME*, Advanced Intelligent Systems, vol. 7, no. 1, 2400304, 2025. DOI: 10.1002/aisy.202400304.
23. C. Bishop, and N Nasrabadi, *Pattern Recognition and Machine Learning*, Springer, New York, pp. 738, 2006.
24. T. Chen, and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Proc. Association for Computing Machinery. pp. 785–794, 2016. DOI:

10.1145/2939672.2939785.
25. C. Bentéjac, A. Csörgő and G. Martínez-Muñoz, *A comparative analysis of gradient boosting algorithms*, Artificial Intelligence Review, vol. 54, no. 3, pp. 1937–1967, 2021. DOI: 10.1007/s10462-020-09896-5.
26. H. Ramchoun et al., *Multilayer Perceptron: Architecture Optimization and Training*, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 4, no. 1, pp. 26–30, 2016. DOI: 10.9781/ijimai.2016.415.
27. W. Wang et al., *Generalized Autoencoder: A Neural Network Framework for Dimensionality Reduction*, Proceedings of CVPR Workshops, pp. 490–497, 2014.
28. S. A. Hicks et al., *On evaluation metrics for medical applications of artificial intelligence*, Scientific Reports, vol. 12, no. 1, 5979, 2022. DOI: 10.1038/s41598-022-09954-8.
29. L. Muflikhah et al., *Feature Selection for Hypertension Risk Prediction Using XGBoost on Single Nucleotide Polymorphism Data*, Healthcare Informatics Research, vol. 31, no. 1, pp. 16–22, 2025. DOI: 10.4258/hir.2025.31.1.16.
30. *Mini-Mental State Examination (MMSE)*, [Internet]. Available from: https://meded.temertymedicine.utoronto.ca/sites/default/files/assets/resource/document/mini-mental-state-examinationmmse.pdf
31. Ž. Đ. Vujovic, *Classification Model Evaluation Metrics*, International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, 2021. DOI: 10.14569/IJACSA.2021.0120670.
32. A. METTAHRI, S. El Khamlichi, M. El Haloui, and B. Ettaki, *Demonstration video of Alzheimer diagnosis tool*, [Internet], 2026. Available from: https://drive.google.com/file/d/1Y8jf_Zpgx0vKa0JZ-LF75OVc8Qd0KdSm/view
33. M. Panpalli Ates et al., *Analysis of genetics and risk factors of Alzheimer's Disease*, Neuroscience, vol. 325, pp. 124–131, 2016. DOI: 10.1016/j.neuroscience.2016.03.051.
34. C. Tremblay et al., *Uncovering atrophy progression pattern and mechanisms in individuals at risk of Alzheimer's disease*, Brain Communications, vol. 7, no. 2, fcaf099, 2025. DOI: 10.1093/braincomms/fcaf099.
35. A. Pacholko and C. Iadecola, *Hypertension, Neurodegeneration, and Cognitive Decline*, Hypertension, vol. 81, no. 5, pp. 991–1007, 2024. DOI: 10.1161/HYPERTENSIONAHA.123.21356.
36. J. Cummings et al., *Diabetes: Risk factor and translational therapeutic implications for Alzheimer's disease*, European Journal of Neuroscience, vol. 56, no. 9, pp. 5727–5757, 2022. DOI: 10.1111/ejn.15619.
37. A. METTAHRI, S. El Khamlichi, M. El Haloui, and B. Ettaki, *An Optimized Machine Learning Model for Alzheimer's Disease Diagnosis*, In Proceedings of Computer Sciences for Sustainable Development (DATA 2025), Tangier, Morocco, pp. 182–192, Springer, 2026. DOI: 10.1007/978-3-032-22567-2_16.
38. T. Chen, and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Proc. Association for Computing Machinery. pp. 785–794, 2016. DOI: 10.1145/2939672.2939785.
39. A. Mohamed, M. Abdelrehim, and R. Al-Barazie, *Context matters in machine learning-based disease prediction with insights from diverse clinical and symptom data*, Scientific Reports, vol. 15, no. 1, 42669, 2025. DOI: 10.1038/s41598-025-26855-8.