STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 2026, pp 0–23.
Published online in International Academic Press (www.IAPress.org)

A Robust Machine Learning Pipeline for Alzheimer's Disease Detection Using Data Balancing and Feature Selection

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 currently considered one of the main global health concerns. Till now, there is no cure for this irreversible brain disorder. However, doctors can benefit from an early AD detection to prescribe an effective treatment that will significantly reduce the progression of the illness. In response to these challenges, the goal of our research is to implement a reliable Machine Learning (ML) model using a publicly available AD dataset. To address the class imbalance issue, several sampling methods are utilized including Tomek Link (T-Link), Random Over-Sampling (ROS), Adaptive Synthetic Sampling (ADASYN), Random Under-Sampling (RUS), the conventional Synthetic Minority Oversampling Technique (SMOTE) algorithm and its hybrid variant SMOTE Edited Nearest Neighbors (SMOTE-ENN). Additionally, we carried out feature selection to determine key attributes for predicting AD by utilizing popular methods, including correlation matrix, Recursive Feature Elimination (RFE), LassoCV, Random Forest (RF) and eXtreme Gradient Boosting (XGBoost) feature importance. Then, we developed predictive models using numerous ML methods: Logistic Regression (LR), Bagging meta-estimator (BaggingME), RF, XGBoost, Decision Tree (DT) and K-Nearest Neighbors (KNN). Model evaluation was performed using repeated nested cross-validation, with preprocessing, feature selection, data balancing and hyperparameter optimization restricted to the training data to prevent information leakage. The results revealed that XGBoost with feature selection and hyperparameter tuning under the data-without-balancing (DWB) configuration provided the best overall discrimination, obtaining 72.72% accuracy, 66.56% precision, 68.33% recall, 67.44% F1-score and 80.31% AUC. The results also showed that data balancing did not consistently improve discrimination, but mainly affected the trade-off between precision and recall.

Keywords Alzheimer's Disease diagnosis, Feature selection, Machine Learning, Sampling techniques

AMS 2010 subject classifications 68T05, 62H30

DOI: 10.19139/soic-2310-5070-4377

1. Introduction

As the predominant kind of dementia, Alzheimer's Disease (AD) leads to progressive memory loss and cognitive decline that severely affect patients and their families [1]. In 2021, over 57 million individuals across the globe suffered from dementia, and roughly 10 million additional instances occurring annually [2]. AD constitutes nearly two-thirds of dementia cases and is influenced by factors such as lifestyle, medical history and environment [3]. Although no cure currently exists, early detection and identification of risk factors can slow or even prevent its progression through therapeutic and lifestyle interventions.

Machine Learning (ML) models have emerged as effective tools in predicting AD. However, like most medical datasets, AD data are often imbalanced, with far more healthy than diseased cases, leading to biased

*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)
Copyright © 2026 International Academic Press

predictions and overfitting. To address this, balancing approaches such as Adaptive Synthetic Sampling (ADASYN) and the conventional Synthetic Minority Oversampling Technique (SMOTE), previously used in COVID-19 mortality prediction [4], [5], can be employed. Moreover, AD datasets frequently contain heterogeneous and high-dimensional variables, complicating analysis.

This study aims to improve AD prediction by systematically evaluating feature selection, data balancing and ML classification under a rigorous validation framework. Additionally, earlier studies often relied on restricted, region-specific datasets focusing mainly on biomarkers or Magnetic Resonance Imaging (MRI) data, which may require specialized acquisition procedures and resources.

In that context, the goal of our study is:

• To apply several ML classifiers including, Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Bagging meta-estimator (BaggingME), Logistic Regression (LR), Decision Tree (DT) and K-Nearest Neighbors (KNN), to identify the most accurate model for AD diagnosis.

• To explore multiple feature selection methods namely, Recursive Feature Elimination (RFE), LassoCV, RF, XGBoost, for identifying key risk factors.

• To test different strategies for data balancing techniques: ADASYN, Random Over-Sampling (ROS), Tomek Link (T-Link), SMOTE Edited Nearest Neighbors (SMOTE-ENN), Random Under-Sampling (RUS) and SMOTE, to evaluate their effects on model.

• To evaluate model performance using repeated nested cross-validation while preventing information leakage during preprocessing, feature selection, balancing and hyperparameter optimization.

The key contribution of this work lies in using a recently published dataset that covers a wide variety of attributes, such as clinical assessments, variables related to daily habits, medical records and demographic data. This is unlike the most previous research, which has mostly concentrated on biomarkers and data from MRI. Furthermore, our approach incorporates numerous types of ML classifiers, combines a variety of techniques for selecting features and balancing data within a systematic comparative framework. The study focuses on the rigorous integration and comparison of these techniques under a common evaluation framework. Furthermore, our study employs hyperparameter optimization and repeated nested cross-validation, with preprocessing, feature selection and data balancing performed within the training data to prevent information leakage and provide a more reliable evaluation. As far as we are aware, prior research has hardly ever integrated all of these components into a single framework.

This study is further organized this way: Section 2 surveys the existing literature; Section 3 introduces the research background; Section 4 outlines the data and methods adopted for this research; Section 5 shares the outcome of the experiments and Section 6 offers an analysis of the obtained results. Section 7 sums up this work.

2. Related works

The application of ML in medical fields, especially for AD diagnosis, has recently gained popularity. To create most reliable and efficient predictive algorithms, a lot of studies have been done. For example, Demir and Selvitopi [6] used a publicly accessible longitudinal MRI data from the OASIS and Kaggle platforms to investigate AD classification. They employed several ML algorithms to predict AD, namely Artifical Neural Networks (ANN), LR, RF, DT and Support Vector Machine (SVM). Overall, RF model outperforms the other models obtaining 86% accuracy, 80% recall, 94% precision, 86% F1-score and 87% AUC. The ANN model performs exceptionally well, with a recall of 91%. Alshamlan et al. [7] investigated SVM classifier using various methods for selecting features, which included minimum Redundancy Maximum Relevance (mRMR) technique, the Correlation-based Feature Selection (CFS) method, F-score, the Chi-Square analysis and Genetic Algorithm (GA) in order to enhance model performance. SVM attained the highest accuracy of 84% when using mRMR with 40 selected genes and F-score with 20 selected genes. Singh et al. [8] employed a public dataset. They suggested a results performance comparison approach of three ML algorithms including LR, RF and Gradient Boosting (GB) to identify the best model. The outcome of this study showed that the GB model outscored the other algorithms obtaining 72% F1-score and 73% recall. Additionally, Kavitha et al. [9] used Chi-Square and correlation coefficient approaches to determine key features from MRI data obtained from the OASIS MRI AD dataset. They used

voting, XGBoost, RF, SVM and DT models; RF exhibited the best accuracy of 86.92%. Likewise to this, Oriol et al. [10] investigated AD classification using genomic data derived from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. Their study compared several classifiers, including Recursive Partitioning and Regression Trees (RPART), RF, KNN, SVM, Bootstrap Stage-Wise Model Selection (BSWiMS) and the ensemble of all these methods. The ensemble algorithm demonstrated better results when measured against the other algorithms, recording 71.9% AUC. Additionally, some studies tried to deal with the issue of imbalanced AD datasets using data balancing techniques, in order to enhance their results and avoid skewed model performance. In this regard, a new Convolutional Neural Network (CNN) model was created by Ahmed et al. [11] from the ground up called Detection of Alzheimer's Disease Network (DAD-Net) to accurately classify the stages of AD. DAD-Net model was built using a multi-class Kaggle MRI image dataset exhibiting a high level of data imbalance. The ADASYN algorithm was used to address the imbalanced dataset problem and enables more accurate AD classification. The findings revealed that proposed DAD-Net model reached a 99.19% F1-Score using ADASYN significantly better than 88.09% F1-Score achieved without using ADASYN. Zadgaonkar et al. [12] developed a RF model with the aim of detecting AD and dementia in advance utilizing lifestyle data collected through the National Health and Ageing Trends Study (NHATS). Since the initial dataset include more than 500 parameters, a Spearman's correlation technique was applied to identify the most relevant lifestyle parameters. RF was trained with the shortlisted parameters and achieved 84% accuracy, 100% recall, 76% precision and 86% F1-score. Mohamed et al. [13] used the same dataset and combined Chi-Square and RFE feature selection with SMOTE to evaluate several ML and deep learning models, with LightGBM achieving the highest accuracy of 76.48%. Similarly, Mahmoud and Salama [14] evaluated multiple classifiers and balancing techniques on the same dataset using a leakage-safe framework. Their best model, AdaBoost, achieved 72.97% accuracy, 73.63% recall, 69.25% F1-score and 80.00% AUC.

The reviewed studies use substantially different data modalities, including MRI [6], [9], [11], genomic data [7], [10] and lifestyle or tabular variables [8], [12–14]. Therefore, their reported performance values are not directly comparable because they are influenced by differences in datasets, feature types, class distributions, preprocessing procedures and validation protocols. MRI and genomic approaches may provide more disease-specific information but generally require specialized acquisition, processing, or laboratory procedures. In contrast, tabular approaches based on demographic, lifestyle, medical-history, and clinical variables use information that is generally more accessible. Accordingly, the studies are summarized in Table 1 according to their data characteristics and evaluation procedures rather than interpreted as a direct performance ranking.

3. Background

3.1. Feature selection techniques

The purpose of feature selection is to identify a subset of predictors that may efficiently define the input data when mitigating the impacts of noise or unrelated factors and still yielding reliable prediction outcomes [15].

There are many feature selection approaches that may be categorized into three primary types: filter approaches, wrapper techniques and embedding strategies.

Filter Methods: Filter approaches evaluate features' discriminative capacity only on the basis of the data's inherent characteristics [16]. They are used to eliminate the less important factors before classification.

Wrapper Methods: Wrapper approaches find the best subset of features by repeatedly training and assessing a ML model [17]. Indeed, a subset of variables that maximizes classification performance or the goal function, can be found using a variety of search algorithms. One of the most common example of wrapper technique is RFE which iteratively eliminates features according to their relevance scores until the needed number of features is obtained [17].

Embedded Methods: Feature selection is done by embedded approaches when training the model. They combine the upsides of both the filter and wrapper techniques. The model may dynamically select the most pertinent features based on the training process since feature selection is included into the model training process [17]. One of the most popular embedding techniques is the regularization method for linear models known as Least Absolute

Table 1. Summary of related studies on ML-based AD prediction

Study	Dataset type	ML approach	Balancing and validation strategy	Reported performance
[6]	MRI-derived and clinical data	RF, DT, LR, SVM, ANN	No balancing reported / train-test evaluation	RF: Accuracy 86%, Precision 94%, Recall 80%, F1-score 86%, AUC 87%
[7]	Genomic data	SVM with mRMR, CFS, F-score, Chi-square and GA feature selection	No balancing reported / 10-fold cross-validation	mRMR/F-score + SVM: Accuracy 84%
[8]	Demographic, genetic and lifestyle tabular data	LR, RF, GB; GridSearchCV	No balancing reported / 80:20 train-test split	GB: Accuracy 73.93%, Precision 72%, Recall 73%, F1-score 72%
[9]	MRI-derived data	DT, RF, SVM, GB, Voting with feature selection	No balancing reported / validation-test evaluation	RF: Accuracy 86.92%, Precision 85%, Recall 81%, F1-score 80%
[10]	Genomic/SNP data	RPART, RF, KNN, SVM and ensemble with genomic feature selection	No balancing reported / cross-validation + independent validation subset	Ensemble: AUC 71.9%
[11]	MRI images	DAD-Net CNN	ADASYN / 80:10:10 train-validation-test split	DAD-Net: Accuracy 99.22%, F1-score 99.19%, AUC 99.91%
[12]	Lifestyle and health-related tabular data	RF, ANN with feature selection	Undersampling / 80:20 train-test split + 5-fold cross-validation	RF: Accuracy 84%, Precision 76%, Recall 100%, F1-score 86%
[13]	Demographic, genetic and lifestyle tabular data	Deep Neural Network (DNN), LR, RF, AdaBoost, XGBoost, GB, LightGBM; Chi-square + RFE	SMOTE / stratified 80:20 train-test split	LightGBM: Accuracy 76.48%; DNN: Accuracy 73%, Precision 72%, Recall 76%, F1-score 74%
[14]	Demographic, genetic and lifestyle tabular data	11 classifiers; RFE, MI, PCA; tuning + voting/stacking	SMOTE, Borderline-SMOTE, ADASYN / 80:20 train-test split + 10-fold stratified cross-validation	AdaBoost: Accuracy 72.97%, Precision 65.37%, Recall 73.63%, F1-score 69.25%, AUC 80.00%

Shrinkage and Selection Operator (LASSO). LASSO imposes an L1 penalty on the coefficients of the regression. Features with non-zero coefficients are seen as important [17].

To quantitatively assess the stability and consistency of the feature-selection results, Jaccard similarity and Spearman rank correlation were also used. Jaccard similarity measures the overlap between two sets of selected features and is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Where A and B two sets of selected features of selected features [18]. In this study, Jaccard similarity was used to evaluate both the stability of the Top-k features across different data partitions for the same feature-selection method and the consistency of the selected feature sets across different methods. Spearman rank correlation was

used to assess the similarity between complete feature rankings and is defined as:

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

Where d_i is the difference between the ranks assigned to feature i in two rankings, and n is the total number of ranked features [19]. In this study, Spearman correlation was used to evaluate ranking stability across data partitions and ranking consistency among the different feature-selection methods.

3.2. Data balancing approaches

To solve the issues brought on by class imbalance that come up when working with medical datasets, advanced methods and frameworks have been developed for dealing with this kind of datasets [20].

ROS is one of the oversampling techniques in which instances to be replicated are selected at random [21].

SMOTE is an advanced oversampling method that tries to improve the boundaries of the less represented class through creating fake instances within that class, instead of simply copying the current examples, to prevent overfitting. [21].

SMOTE-ENN it merges the data filtering capability of the ENN method with the artificial sample generation power of SMOTE technique. While ENN eliminates examples from the dominant class that are mistakenly categorized by their closest neighbors, SMOTE creates synthetic instances. This improves the performance of the classifier by lowering noise and the possibility of overfitting [22].

ADASYN operates as a data-driven technique which creates artificial cases for underrepresented class instances presenting greater classification difficulty. Its objective is to improve the model's capacity to generalize on more complex examples in order to increase performance, particularly with datasets with difficult decision boundaries [22]. Indeed, ADASYN can adaptively synthesize a subset of the examples depending on the distribution and automatically calculates how many synthetic examples needed for each minority class instance.

RUS when a random selection of the majority class samples is eliminated.

T-Link is viewed as an improvement over the Nearest-Neighbor Rule (NNR). T-Link specifies boundary pairings in a data distribution. When two instances, a and b, have class (a) class (b) and are closest to one other, they are referred to as a T-Link pair [21]. Once Tomek Links are detected, cases belonging to the positive class are retained in the dataset while those in the majority class negative category are removed [23].

3.3. ML algorithms

The following is a description of the ML algorithms we employed:

DT is a supervised learning method that is mainly applied for classification issues, while it can also be used for regression problems. It is made up of leaf nodes that show final predictions and inner nodes that represent decision points and branches that suggest potential outcomes. To make effective decisions, the program analyzes the dataset for trends [24].

LR represents a a popular ML algorithm used for classification purposes. Using past data, it forecasts the likelihood that an instance will fall into a particular class. Furthermore, LR establishes the connection involving a number of independent variables and a outcome variable [25].

KNN is a supervised ML technique that can be applied to classification and regression problems. The K closest observations, or neighbors, within the attribute space are taken into account for making predictions. The new observation, identified by KNN, corresponds to the dominant class among these K neighboring instances [24].

RF is an ensemble of classifiers made up of decision trees that are produced with two different randomization sources. Initially, a random sample of the original data, equal in size to the training set is used to train each decision tree separately. A second source of randomization in random forests is attribute sampling. To determine the best split, a group of input variables is randomly selected for every node split. The ensemble's final prediction is determined by majority vote [26].

XGBoost is an extremely scalable collection of decision trees which operates through the gradient boosting methodology. It employs the gradient boosting technique, where multiple decision trees are generated sequentially,

fixing errors of the earlier trees. Similar to gradient boosting, XGBoost aims to reduce a loss function to produce an additional enhancement of the objective function [26].

BaggingME functions as an ensemble approach that may be utilized for issues involving classification (BaggingClassifier) and regression (BaggingRegressor). It makes predictions using the standard bagging technique. The ultimate prediction is formed by merging the outcomes from numerous base estimator instances that were trained on various random segments from the initial collection of data [27].

3.4. Evaluation metrics

For binary classification tasks, the confusion matrix provides four fundamental quantities: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). These values were used to derive the following metrics for evaluating our proposed models:

Accuracy represents the percentage of all instances in the evaluation that were correctly classified. It can be calculated using the equation (3) [28].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Precision determines the rate of properly classified positive cases among all cases predicted to be positive. It assesses the percentage of recovered samples that are actually pertinent. It is generally formulated as following: (4) [28].

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Recall refers to the ratio of all instances assigned to the positive class to correctly classified positive observations. It indicates the fraction of properly identified positive instances and is sometimes called the sensitivity. It can be computed using equation (5) [28].

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

F1-Score it balances recall together with precision score. This metric is defined by the formula (6) [28].

$$F1_{Score} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6}$$

Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC) are commonly used to assess classifier performance. The ROC curve provides a graphical representation used to evaluate the performance of a classification algorithm over the entire range of decision thresholds. It illustrates how the False Positive Rate (FPR) and the True Positive Rate (TPR) are traded off. The TPR, also known as recall or sensitivity, and the FPR are defined by (7) and (8) [29].

$$TPR = Recall \tag{7}$$

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

The AUC provides a single scalar value summarizing the overall discriminative ability of the classifier. An AUC value of 0 indicates extremely poor classification performance, whereas an AUC value of 1 denotes ideal classification. Higher AUC values, therefore reflect better model performance [29].

3.5. Calibration analysis

Calibration analysis examines how closely a model's predicted probabilities reflect the outcomes actually observed in the data [30]. Good calibration indicates that the estimated probabilities are consistent with the corresponding

observed event frequencies. It can be evaluated using calibration curves, calibration intercept, slope and the Brier score. The Brier score provides a measure of probabilistic prediction accuracy by quantifying the squared difference between predicted probabilities and observed binary outcomes [31] and is defined as:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{9}$$

Where N is the number of observations, p_i is the predicted probability for observation i, and y_i is the corresponding observed outcome (0 or 1). Lower Brier scores indicate better probabilistic prediction accuracy, with 0 representing perfect predictions.

3.6. Decision Curve Analysis

Decision Curve Analysis (DCA) is a decision-analytic method for evaluating the potential utility of prediction models by quantifying their net benefit across a range of threshold probabilities [32]. Net benefit combines true-positive decisions with a penalty for false-positive decisions according to the selected threshold probability. DCA compares model-guided decisions with the default strategies of Treat All, where all individuals are flagged for further assessment, and Treat None, where no individuals are flagged [32]. In this context, the threshold probability represents the predicted risk at which an individual would be flagged for further assessment.

4. Data and methods

4.1. Data description

In our study we employed a public AD datasets obtained from Kaggle platform. This dataset [33] is composed of 74,283 patients from 20 countries. It offers information on risk factors for AD and consists of 25 independent features. It encompasses genetic, medical, lifestyle and demographic factors. The characteristics age, BMI, cognitive test score, and education level are numerical, while the remaining features are categorical, such as smoking status, gender and Alzheimer's diagnosis. The target variable of our investigation is Alzheimer's diagnosis, which is represented by two output classes (Yes, No). Furthermore, there are 30,713 observations classified as Yes and 43,570 cases labeled as No in the dataset. The dataset was examined for missing values, and no missing values were identified; therefore, no imputation was required. Table 2 provides additional details about the dataset. The AD dataset is imbalanced. This may result in biased models that favor the dominant class and struggle to properly learn the characteristics of the underrepresented class.

4.2. Proposed solution

The design of our proposed solution was guided by the the aim of obtaining a high-performing ML model which can serve as an ML framework that may support AD screening and clinical decision-making. Achieving this objective requires properly addressing the issue of dataset imbalance. Model evaluation was performed using repeated nested cross-validation. The outer stratified 5-fold cross-validation was repeated five times, resulting in 25 outer evaluations. Within each outer training partition, the inner 5-fold cross-validation was used for model configuration and hyperparameter optimization, while the outer test fold remained completely untouched until the final performance evaluation. The same outer partitions were used across the compared methods to ensure consistent comparisons. The overall procedure is illustrated in Figure 1. After dataset collection, we started with the data preprocessing step. To prevent information leakage, preprocessing was performed within the cross-validation procedure. All preprocessing parameters were fitted exclusively on the training data and then applied to the corresponding validation or test data. Similarly, feature selection, data balancing and hyperparameter optimization were performed using only the training data within the nested cross-validation procedure. Numerical features were processed separately from categorical features. Numerical features were standardized using StandardScaler, with the mean and standard deviation estimated exclusively from the corresponding training data and then applied to the

Table 2. AD 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

validation and test data. Categorical features were encoded using one-hot encoding, with the encoder fitted only on the training data and subsequently applied to the validation and test data. For feature selection, we employed several approaches, including RFE, LassoCV, RF feature importance and XGBoost feature importance, to identify the most relevant attributes. Feature selection was first analyzed to compare the different methods and assess the stability and consistency of the identified features across data partitions. The feature rankings obtained from this analysis were not directly used for predictive modelling. During predictive modelling, feature selection was performed again within the training procedure, where the feature-selection method and the number of selected features were determined using inner validation without access to the held-out test data. Stability across training partitions was assessed using Spearman rank correlation for the complete feature rankings and Jaccard similarity for the Top-k feature sets. Consistency among the different feature-selection methods was also evaluated using Spearman rank correlation and Jaccard similarity. Furthermore, data balancing techniques, namely ROS, SMOTE, SMOTE-ENN, ADASYN, RUS and T-Link were used to address class imbalance, together with data without balancing (DWB) as the reference condition. All data-balancing procedures were applied exclusively to the training data within the cross-validation process. After that, we applied the most widely employed ML algorithms: XGBoost, LR, KNN, DT, BaggingME and RF using the repeated nested cross-validation procedure AD classification (Figure 1).

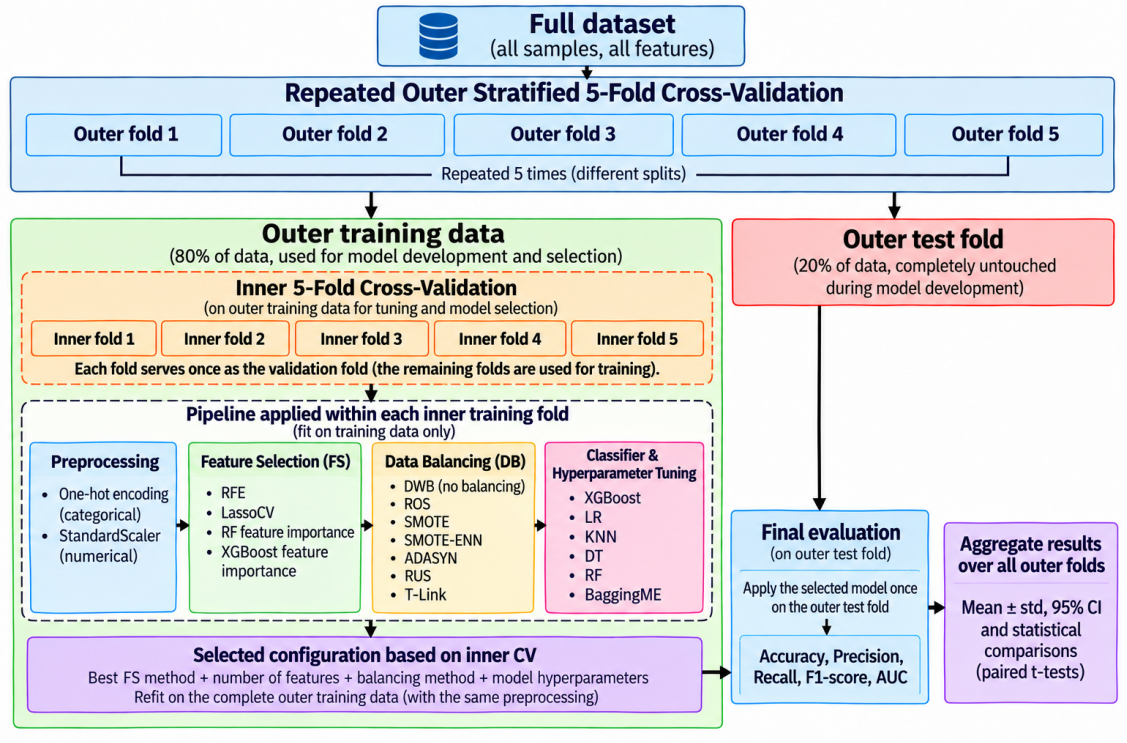


Figure 1. Overview of the proposed solution.

5. Results of experiments

This section exhibits the results of our experiments.

5.1. Feature selection analysis

First, we performed feature selection on the AD dataset to identify the most relevant attributes employing several approaches, including RFE, LassoCV, RF feature importance and XGBoost feature importance. For each method, the mean rank and score and their corresponding standard deviations were calculated across the 25 outer-training partitions. and the features were ordered according to their mean rank. Table 3 and Table 4 present the Top-10 features identified using RFE and LassoCV, respectively. The RFE and LassoCV results show substantial agreement among the highest-ranked features. Country, Age, GeneticRiskFactorAPOE_ε4_allele and FamilyHistoryOfAlzheimer were the four highest-ranked features in both methods, although their ranking order differed. IncomeLevel and AlcoholConsumption also appeared among the highly ranked features. The RF and XGBoost feature-importance results are presented in Figure 2. RF assigned the highest importance to Age, followed by Country, BMI, CognitiveTestScore, EducationLevel and GeneticRiskFactorAPOE_ε4_allele. In contrast, XGBoost assigned the highest importance to GeneticRiskFactorAPOE_ε4_allele, followed by Country, Age and FamilyHistoryOfAlzheimer. Overall, the results show that several features, particularly Age, Country and GeneticRiskFactorAPOE_ε4_allele, were highly ranked across multiple methods, while the relative importance of the remaining features varied according to the feature-selection method.

For predictive modelling, feature selection was performed within the training partitions to avoid information leakage. To provide a more rigorous assessment of feature selection, the stability of each feature-selection method

Table 3. Top 10 features identified using RFE

Feature	Rank	Rank (mean ± std)	Score (mean ± std)
Country	1	1.00 ± 0.000	24.00 ± 0.000
Age	2	2.00 ± 0.000	23.00 ± 0.000
GeneticRiskFactorAPOE_ε4_allele	3	3.00 ± 0.000	22.00 ± 0.000
FamilyHistoryOfAlzheimer	4	4.00 ± 0.000	21.00 ± 0.000
IncomeLevel	5	7.88 ± 2.728	17.12 ± 2.728
AlcoholConsumption	6	8.84 ± 4.180	16.16 ± 4.180
AirPollutionExposure	7	9.68 ± 4.130	15.32 ± 4.130
UrbanVSRuralLiving	8	10.00 ± 4.592	15.00 ± 4.592
DietaryHabits	9	10.52 ± 3.874	14.48 ± 3.874
CholesterolLevel	10	11.04 ± 4.306	13.96 ± 4.306

Table 4. Top 10 features identified using LassoCV

Feature	Rank	Rank (mean ± std)	Coefficient (mean ± std)
GeneticRiskFactorAPOE_ε4_allele	1	1.00 ± 0.000	0.236338 ± 0.002
Country	2	2.08 ± 0.277	0.213783 ± 0.006
Age	3	2.92 ± 0.277	0.206358 ± 0.001
FamilyHistoryOfAlzheimer	4	4.00 ± 0.000	0.151880 ± 0.001
IncomeLevel	5	7.84 ± 2.461	0.003574 ± 0.002
UrbanVSRuralLiving	6	8.70 ± 4.036	0.003500 ± 0.002
AlcoholConsumption	7	9.48 ± 4.273	0.002909 ± 0.001
AirPollutionExposure	8	10.28 ± 4.458	0.002669 ± 0.002
EmploymentStatus	9	10.84 ± 3.782	0.002279 ± 0.001
DietaryHabits	10	11.22 ± 4.183	0.002395 ± 0.002

across training partitions was evaluated using Spearman rank correlation for the complete feature rankings and Jaccard similarity for the Top-k feature sets. The results are presented in Table 5.

Table 5. Stability of feature-selection methods across training partitions

Method	Spearman (mean ± std)	Jaccard Top-5 (mean ± std)	Jaccard Top-6 (mean ± std)	Jaccard Top-8 (mean ± std)	Jaccard Top-10 (mean ± std)
LassoCV	0.683 ± 0.105	0.701 ± 0.102	0.594 ± 0.129	0.546 ± 0.128	0.585 ± 0.121
RFE	0.762 ± 0.083	0.713 ± 0.116	0.592 ± 0.119	0.558 ± 0.126	0.564 ± 0.111
RF Importance	0.914 ± 0.035	1.000 ± 0.000	1.000 ± 0.000	0.694 ± 0.103	0.647 ± 0.115
XGBoost Importance	0.878 ± 0.036	0.710 ± 0.112	0.590 ± 0.119	0.535 ± 0.120	0.570 ± 0.110

The stability analysis shows that RF importance was the most stable method across training partitions, with the highest mean Spearman correlation (0.914 ± 0.035) and identical Top-5 and Top-6 feature sets across all comparisons (Jaccard = 1.000 ± 0.000). XGBoost importance also showed high complete-ranking stability (0.878 ± 0.036), followed by RFE (0.762 ± 0.083) and LassoCV (0.683 ± 0.105). Overall, the results indicate that the feature rankings showed varying degrees of reproducibility across different training partitions, with RF and XGBoost importance exhibiting greater stability than RFE and LassoCV. Consistency among the different feature-selection methods was further examined using Spearman rank correlation and Jaccard similarity. The results are presented in Table 6.

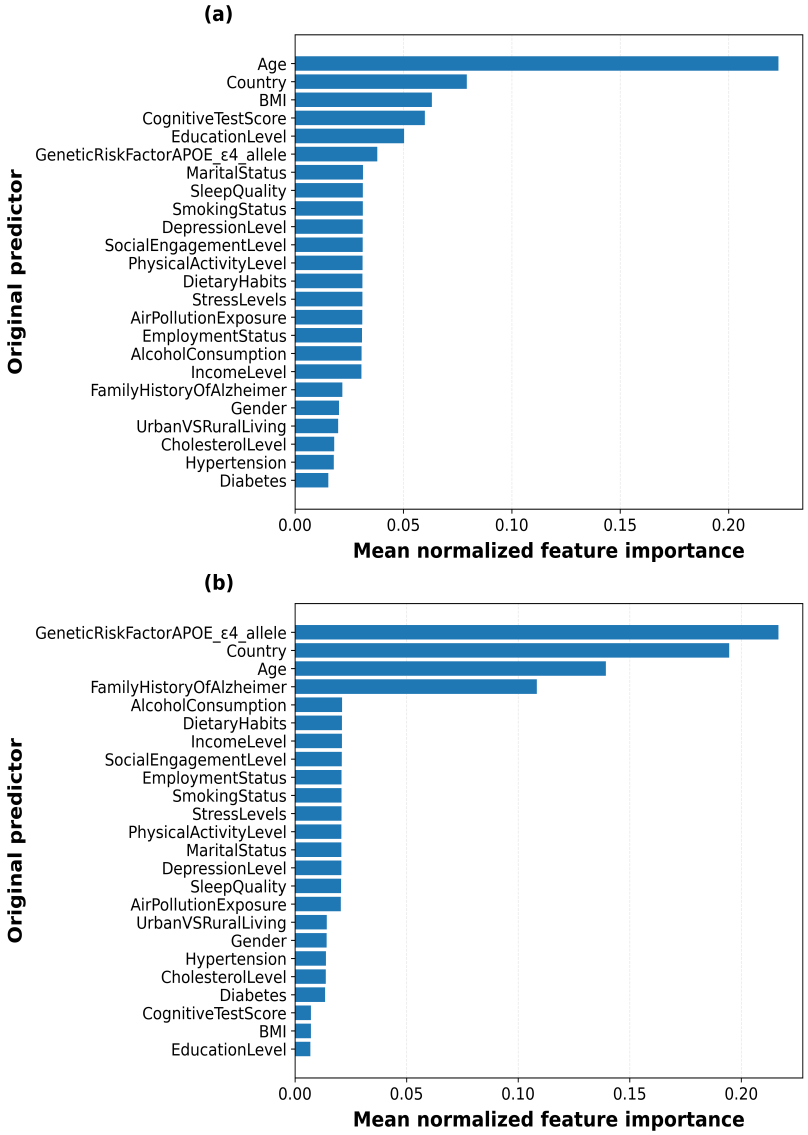


Figure 2. Feature importance according to (a) RF and (b) XGBoost.

Table 6. Consistency among feature-selection methods

Method comparison	**Jaccard Top-5 (mean ± std)**	**Jaccard Top-6 (mean ± std)**	**Jaccard Top-8 (mean ± std)**	**Jaccard Top-10 (mean ± std)**	**Spearman ρ (mean ± std)**
LassoCV vs RF_Importance	0.250 ± 0.000	0.333 ± 0.000	0.239 ± 0.028	0.218 ± 0.050	0.080 ± 0.084
LassoCV vs XGB_Importance	0.667 ± 0.000	0.577 ± 0.105	0.484 ± 0.113	0.474 ± 0.068	0.634 ± 0.076
RFE vs LassoCV	0.920 ± 0.145	0.886 ± 0.143	0.895 ± 0.126	0.935 ± 0.089	0.953 ± 0.017
RFE vs RF_Importance	0.250 ± 0.000	0.333 ± 0.000	0.235 ± 0.021	0.218 ± 0.050	-0.011 ± 0.060
RFE vs XGB_Importance	0.693 ± 0.092	0.586 ± 0.107	0.467 ± 0.111	0.469 ± 0.068	0.703 ± 0.075
RF_Importance vs XGB_Importance	0.250 ± 0.000	0.333 ± 0.000	0.289 ± 0.062	0.332 ± 0.086	0.237 ± 0.040

The consistency analysis shows that the strongest agreement was between RFE and LassoCV, with a mean Spearman correlation of 0.953 ± 0.017 and high Jaccard similarities across all evaluated Top-k feature sets. RFE also showed relatively strong ranking agreement with XGBoost importance (0.703 ± 0.075), followed by LassoCV and XGBoost importance (0.634 ± 0.076). In contrast, RF importance showed relatively low agreement with the other methods despite its high within-method stability. Thus, a method can produce highly stable rankings across training partitions without necessarily identifying the same feature ordering as other selection methods.

Country appeared among the highly ranked features in several feature-selection methods. Since Country is a non-clinical variable that may capture geographic or other population-related differences, we further examined its influence through an ablation analysis. The analysis was performed using the XGBoost classifier under the DWB setting. The same predictive configuration was evaluated on identical outer test partitions with and without the Country feature. Performance was summarized in Table 7 as mean ± std across the five repeated outer cross-validation runs.

Table 7. Performance comparison with and without Country

Configuration	**Accuracy (mean ± std, %)**	**Precision (mean ± std, %)**	**Recall (mean ± std, %)**	**F1-score (mean ± std, %)**	**AUC (mean ± std, %)**
With Country	72.72 ± 0.05	66.56 ± 0.08	68.33 ± 0.08	67.44 ± 0.05	80.31 ± 0.04
Without Country	72.65 ± 0.04	65.27 ± 0.04	72.32 ± 0.20	68.62 ± 0.08	79.39 ± 0.02

As shown in Table 7, removing Country resulted in only a small decrease in accuracy, from 72.72% to 72.65%, while AUC decreased from 80.31% to 79.39% and precision from 66.56% to 65.27%. In contrast, recall increased from 68.33% to 72.32% and F1-score from 67.44% to 68.62%. Overall, excluding Country produced a limited reduction in discrimination while improving the detection of positive cases.

5.2. Model performance comparison

Based on this analysis, feature selection was then incorporated into the predictive modelling procedure, where the selected feature subset was determined within the training partitions to avoid information leakage. We developed six ML algorithms: LR, KNN, DT, RF, XGBoost and BaggingME. We optimized each classifier within the inner cross-validation procedure, using RandomizedSearchCV technique from Scikit-learn Python library (using the range values in Table 8) and evaluated them with different data balancing techniques, including ROS, SMOTE, SMOTE-ENN, ADASYN, RUS and T-Link, in addition to DWB. To assess model performance, we considered Accuracy, Recall, Precision, F1-score together with the AUC score. Model performance was evaluated using repeated nested stratified cross-validation. The outer 5-fold cross-validation was repeated five times, while an inner 5-fold cross-validation was used for feature-selection configuration and hyperparameter optimization. The same outer partitions were maintained across the compared experimental conditions to enable consistent comparisons.

5.2.1. DWB performance :

We first evaluated the classifiers using the original class distribution without data balancing (DWB). This provided a common reference for determining whether the application of balancing techniques produced meaningful changes in predictive performance. Performance was summarized across the repeated outer evaluations using mean ± SD and 95% confidence intervals. The DWB results are presented in Table 9. As shown in Table 9, XGBoost achieved the highest overall performance under DWB, with an accuracy of 72.72% and an AUC of 80.31%. RF obtained the second-highest AUC (79.96%), while DT achieved the highest recall (71.75%). Overall, the results indicate differences in the performance trade-offs among the classifiers, with XGBoost providing the strongest discrimination under the DWB configuration.

The ROC curves of the six classifiers under the DWB configuration are presented in Figure 3, providing a visual comparison of their discriminative performance. The ROC curves further confirm the relatively similar discrimination of the classifiers under DWB, with XGBoost achieving the highest mean AUC (0.803), followed by RF (0.800). Overall, only limited differences were observed among the best-performing classifiers.

Table 8. Hyperparameter search spaces used for model optimization

Classifier	Hyperparameter	Candidate values
LR	Solver	liblinear
	C	0.01, 0.1, 0.5, 1, 2, 5, 10
	Penalty	l1, l2
KNN	n_neighbors	3, 5, 7, 9, 11, 15, 21
	Weights	uniform, distance
	p	1, 2
DT	max_depth	None, 3, 5, 7, 10, 15, 20
	min_samples_split	2, 5, 10, 20
	min_samples_leaf	1, 2, 5, 10
	Criterion	gini, entropy
RF	n_estimators	200, 300, 500
	max_depth	None, 5, 10, 15, 20
	min_samples_split	2, 5, 10
	min_samples_leaf	1, 2, 5
	max_features	sqrt, log2, None
XGBoost	n_estimators	200, 300, 500, 700
	learning_rate	0.01, 0.03, 0.05, 0.1
	max_depth	3, 5, 7, 10, 12, 15, 20
	min_child_weight	1, 3, 5, 10
	subsample	0.6, 0.8, 1.0
	colsample_bytree	0.6, 0.8, 1.0
	gamma	0, 0.1, 0.3, 0.5
	reg_alpha	0, 0.01, 0.1, 0.5, 1.0
	reg_lambda	0.5, 1, 2, 5, 10
BaggingME	n_estimators	50, 100, 200, 300
	max_samples	0.6, 0.8, 1.0
	max_features	0.6, 0.8, 1.0

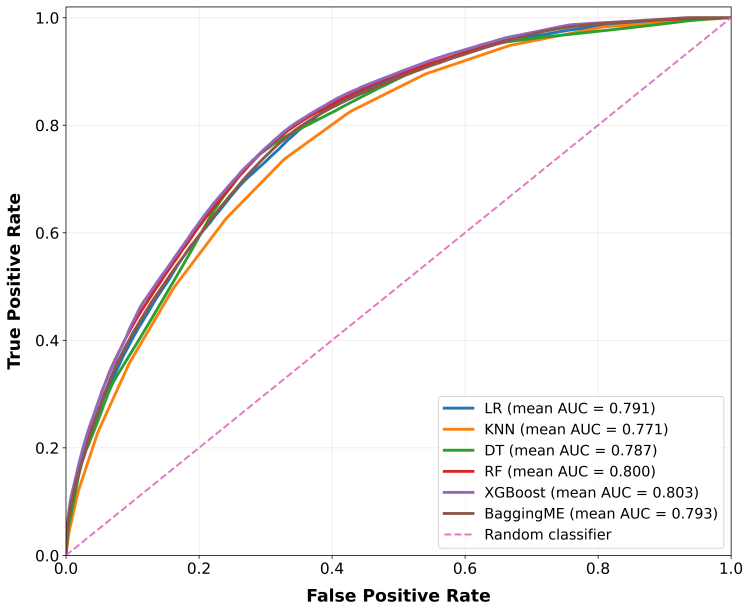


Figure 3. Mean ROC curves of the ML classifiers under the DWB configuration.

Table 9. Performance of the evaluated classifiers under the DWB setting

Model	Accuracy (mean ± std, %)	Precision (mean ± std, %)	Recall (mean ± std, %)	F1-score (mean ± std, %)	AUC (mean ± std, %)	Most frequently selected hyperparameters	
LR	71.53 ± 0.01	67.19 ± 0.02	60.90 ± 0.05	63.89 ± 0.03	79.06 ± 0.00	Solver	liblinear
						C	0.5
						penalty	l2
KNN	70.43 ± 0.10	64.73 ± 0.10	62.61 ± 0.38	63.65 ± 0.20	77.05 ± 0.07	n_neighbors	11
						p	2
						weights	uniform
DT	72.51 ± 0.04	65.24 ± 0.12	71.75 ± 0.21	68.34 ± 0.03	78.68 ± 0.04	criterion	entropy
						max_depth	10
						min_samples_leaf	5
						min_samples_split	10
RF	72.42 ± 0.13	66.77 ± 0.12	66.28 ± 0.38	66.52 ± 0.23	79.96 ± 0.02	max_depth	20
						max_features	sqrt
						min_samples_leaf	5
						min_samples_split	2
						n_estimators	500
XGBoost	72.72 ± 0.05	66.56 ± 0.08	68.33 ± 0.08	67.44 ± 0.05	80.31 ± 0.04	colsample_bytree	0.6
						gamma	0.1
						learning_rate	0.1
						max_depth	5
						min_child_weight	3
						n_estimators	500
						reg_alpha	0.01
						reg_lambda	0.5
						subsample	1.0
BaggingME	71.83 ± 0.06	65.94 ± 0.12	65.92 ± 0.61	65.93 ± 0.25	79.29 ± 0.06	max_features	1.0
						max_samples	0.6
						n_estimators	200

5.2.2. Effect of data balancing :

We then evaluated different data balancing techniques, including ROS, SMOTE, SMOTEENN, ADASYN, RUS and T-Link. DWB was retained as the reference condition. Figure 4 summarizes the mean ROC-AUC obtained by each classifier under DWB and the six balancing techniques, providing a direct comparison of the effect of each balancing strategy across classifiers. The effect of data balancing varied across classifiers. For LR and XGBoost, the AUC values remained very similar across most configurations, while SMOTE-ENN generally produced lower AUC values. DWB achieved the highest mean AUC for DT (78.68%) and XGBoost (80.31%), whereas only small AUC improvements over DWB were observed for some other classifiers, such as RF and BaggingME with T-Link. Overall, data balancing did not consistently improve discrimination compared with DWB.

5.2.3. Best-performing balancing configurations :



Figure 4. AUC performance of ML classifiers across data balancing techniques.

Finally, Table 10 summarizes, for each classifier, the balancing condition that achieved the highest unrounded mean AUC among DWB, ROS, RUS, SMOTE, ADASYN, SMOTE-ENN and T-Link. In addition to AUC, Accuracy, Precision, Recall and F1-score are reported to provide a more complete comparison of classifier performance under the identified best-performing configuration.

Table 10. Performance of classifiers under the best-performing balancing configurations

Model	Selected balancing technique	Accuracy (mean ± std, %)	Precision (mean ± std, %)	Recall (mean ± std, %)	F1-score (mean ± std, %)	AUC (mean ± std, %)
LR	SMOTE	71.53 ± 0.02	63.78 ± 0.03	72.06 ± 0.09	67.67 ± 0.03	79.07 ± 0.01
KNN	T-Link	70.55 ± 0.34	64.51 ± 0.45	63.97 ± 0.44	64.24 ± 0.40	77.15 ± 0.73
DT	DWB	72.51 ± 0.04	65.24 ± 0.12	71.75 ± 0.21	68.34 ± 0.03	78.68 ± 0.04
RF	T-Link	72.50 ± 0.04	64.41 ± 0.04	74.84 ± 0.10	69.23 ± 0.06	80.00 ± 0.02
XGBoost	DWB	72.72 ± 0.05	66.56 ± 0.08	68.33 ± 0.08	67.44 ± 0.05	80.31 ± 0.04
BaggingME	T-Link	71.94 ± 0.04	63.79 ± 0.04	74.34 ± 0.26	68.66 ± 0.10	79.37 ± 0.03

The best-performing balancing configuration differed across classifiers. SMOTE was selected for LR, while T-Link was selected for KNN, RF and BaggingME. In contrast, DWB remained the best configuration for DT and XGBoost, with XGBoost achieving the highest overall AUC 80.31%. These results further indicate that no single balancing technique consistently provided the best performance across classifiers.

5.3. Significance test of model performance differences

To complement the performance comparison, statistical analyses were conducted to determine whether the observed differences were consistent across the repeated outer evaluations. The analysis considered both differences among classifiers and differences between the data-balancing methods and DWB.

Statistical comparisons were first performed to assess differences in AUC among the classifiers under the DWB configuration, using the same outer validation partitions. XGBoost, which achieved the highest mean AUC of 80.31% under DWB, was used as the reference classifier and compared with the other classifiers. Paired two-sided t-tests were applied to the matched repeat-level AUC values obtained from the five repeated outer cross-validation runs. The AUC results, their 95% confidence intervals, and the corresponding p-values are reported in Table 11.

Table 11. AUC performance and statistical comparison of classifiers

Model	**AUC (mean ± std, %)**	**AUC 95% CI (%)**	***p*-value vs best AUC**
XGBoost	80.31 ± 0.04	[80.27, 80.35]	–
RF	79.96 ± 0.02	[79.94, 79.99]	0.00004174
DT	78.68 ± 0.04	[78.62, 78.73]	0.00000023
BaggingME	79.29 ± 0.06	[79.22, 79.37]	0.00000580
LR	79.06 ± 0.00	[79.06, 79.06]	0.00000015
KNN	77.05 ± 0.07	[76.97, 77.13]	0.00000005

The statistical comparisons indicate that XGBoost achieved a significantly higher AUC than all other classifiers under DWB. These results support the selection of XGBoost as the reference classifier for the subsequent analyses.

Given the moderate class imbalance, the effect of the selected best-performing balancing technique for each classifier was further evaluated against DWB using identical validation partitions. AUC performance was summarized as mean ± std with 95% confidence intervals, and paired two-sided t-tests were used to compare the selected balancing technique with DWB. The results are depicted in Table 12.

Table 12. Paired statistical comparison of the selected balancing techniques against DWB

Model	**Selected balancing technique**	**AUC DWB (mean ± std, %)**	**AUC selected balancing technique (mean ± std, %)**	**AUC 95% CI (%)**	***p*-value vs AUC DWB**
LR	SMOTE	79.06 ± 0.00	79.07 ± 0.01	[79.06, 79.07]	0.377033
KNN	T-Link	77.05 ± 0.07	77.15 ± 0.73	[76.24, 78.06]	0.777501
DT	DWB	78.68 ± 0.04	–	–	–
RF	T-Link	79.96 ± 0.02	80.00 ± 0.02	[79.97, 80.02]	0.009821
XGBoost	DWB	80.31 ± 0.04	–	–	–
BaggingME	T-Link	79.29 ± 0.06	79.37 ± 0.03	[79.33, 79.41]	0.003634

Significant AUC improvements over DWB were observed only for RF and BaggingME with T-Link, whereas the differences for LR and KNN were not significant. For DT and XGBoost, DWB remained the best-performing configuration.

5.4. Calibration analysis results

Calibration analysis was performed to evaluate the agreement between predicted probabilities and observed outcomes. Figure 5 presents the calibration curves of the six classifiers under the DWB configuration using the out-of-fold predictions obtained from the repeated nested cross-validation procedure.

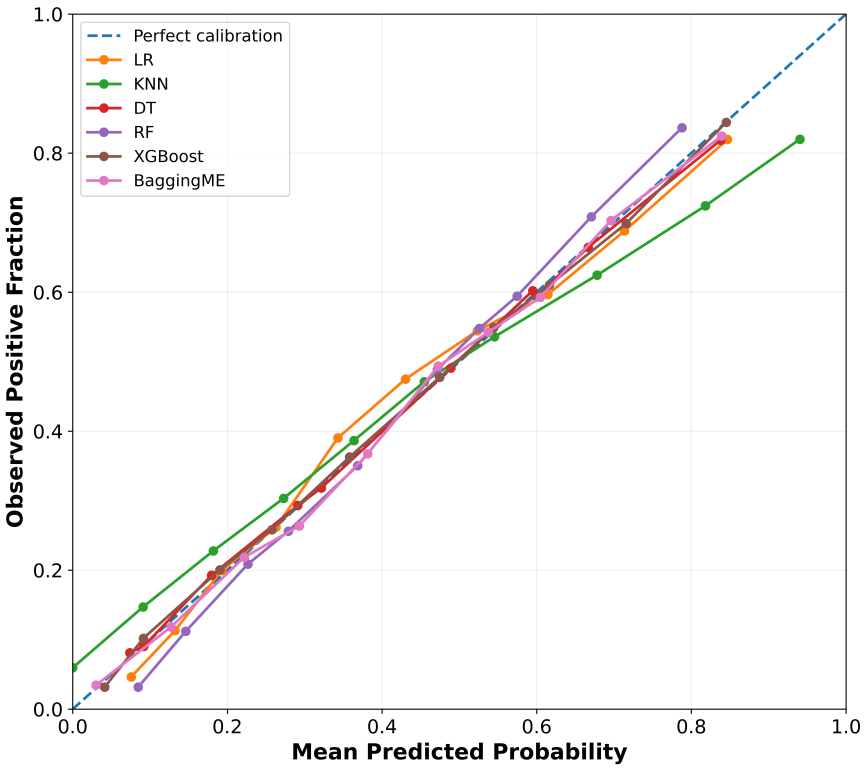


Figure 5. Calibration curves of the six ML classifiers under the DWB configuration based on repeated out-of-fold predictions. The dashed diagonal represents perfect calibration.

Most classifiers followed the perfect-calibration line reasonably closely across much of the probability range. However, some deviations were observed, particularly for KNN, while XGBoost, LR, and BaggingME generally remained closer to the reference line. To provide a quantitative assessment of calibration, the Brier score, calibration intercept, and calibration slope were calculated for each classifier across the five repeated outer cross-validation runs. The corresponding mean ± std values are reported in Table 13, where lower Brier scores, an intercept close to 0, and a slope close to 1 indicate more favorable calibration.

XGBoost achieved the lowest Brier score (0.1770 ± 0.0001), with an intercept close to 0 (-0.0017 ± 0.0002) and a slope close to 1 (0.9914 ± 0.0015). LR and BaggingME also showed slopes close to 1, whereas KNN exhibited the largest deviation from ideal calibration. Overall, XGBoost showed the most favorable combination of discrimination and calibration among the evaluated classifiers.

5.5. DCA results

To further assess the potential decision-analytic utility of data balancing, DCA was performed for XGBoost under DWB and the six balancing configurations. Net benefit was evaluated across threshold probabilities from 0.05 to 0.95, with the Treat All and Treat None strategies included as reference approaches. Figure 6 (a) presents the complete decision curves, while Figure 6 (b) provides an enlarged view of the model net-benefit region to facilitate comparison among the configurations. As shown in Figure 6, most XGBoost configurations provided positive net

Table 13. Calibration performance of the classifiers under the DWB configuration

Model	Brier score (mean ± std)	Calibration intercept (mean ± std)	Calibration slope (mean ± std)
LR	0.1832 ± 0.0000	-0.0005 ± 0.0001	0.9981 ± 0.0001
KNN	0.1933 ± 0.0004	-0.1164 ± 0.0038	0.3004 ± 0.0094
DT	0.1825 ± 0.0002	-0.0156 ± 0.0019	0.9331 ± 0.0062
RF	0.1793 ± 0.0001	0.0570 ± 0.0024	1.2424 ± 0.0078
XGBoost	0.1770 ± 0.0001	-0.0017 ± 0.0002	0.9914 ± 0.0015
BaggingME	0.1814 ± 0.0003	-0.0253 ± 0.0050	1.0014 ± 0.0271

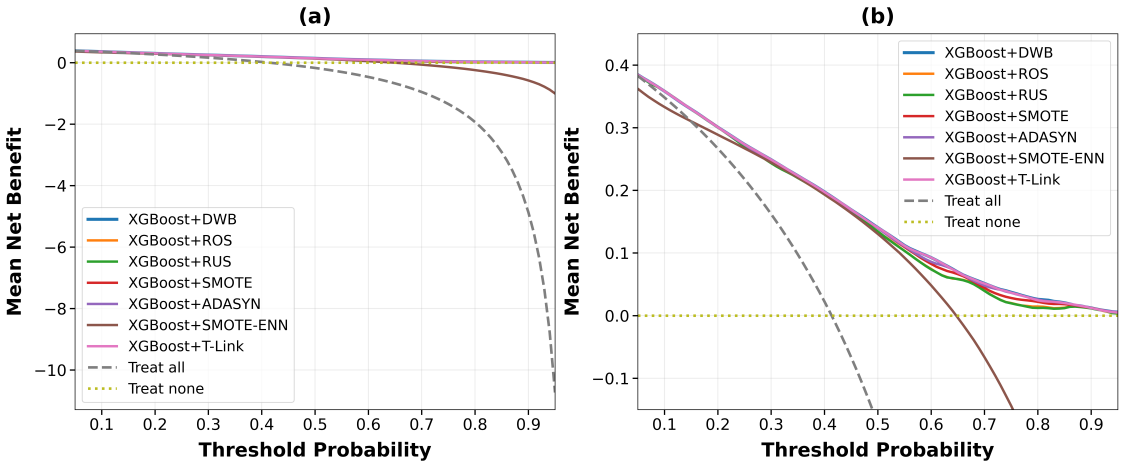


Figure 6. DCA of XGBoost under DWB and different data-balancing configurations: (a) complete decision curves across threshold probabilities from 0.05 to 0.95 and (b) enlarged view of the model net-benefit region.

benefit across a broad range of threshold probabilities and generally outperformed the Treat All and Treat None strategies. The differences among DWB, ROS, RUS, SMOTE, ADASYN and T-Link were relatively small over much of the threshold range, indicating no consistent decision-analytic advantage from balancing. In contrast, SMOTE-ENN showed a marked reduction in net benefit at higher thresholds. These results further indicate that the effect of data balancing depends on the selected decision threshold and does not consistently improve the potential utility of XGBoost.

5.6. Subgroup analysis

Subgroup analysis was performed to examine performance variations across selected patient characteristics: Age and Gender (demographic), EducationLevel (educational), FamilyHistoryOfAlzheimer and GeneticRiskFactorAPOE_ε4_allele (AD-related risk), and Hypertension and Diabetes (health conditions). These variables were selected based on their relevance for subgroup evaluation rather than on the observed model performance. The results are presented in Table 14. The analysis used out-of-fold predictions, while Country was evaluated separately through ablation analysis. Numerical variables were grouped using quantile-based intervals, whereas categorical variables retained their original categories. Subgroup performance was evaluated for groups with adequate sample sizes. Prevalence represents the percentage of AD-positive cases within each subgroup.

Table 14. Performance analysis across patient subgroups

Subgroup variable	Subgroup	Number	Prevalence (%)	Accuracy (mean ± std, %)	Precision (mean ± std, %)	Recall (mean ± std, %)	F1-score (mean ± std, %)	AUC (mean ± std, %)
Gender	Female	37249	41.28	72.73 ± 0.11	66.62 ± 0.15	68.05 ± 0.14	67.32 ± 0.13	80.31 ± 0.04
	Male	37034	41.41	72.70 ± 0.03	66.51 ± 0.03	68.62 ± 0.11	67.55 ± 0.06	80.31 ± 0.03
Age	Q1 [50, 61]	19833	15.65	84.37 ± 0.03	50.89 ± 1.12	4.70 ± 0.18	8.61 ± 0.30	75.57 ± 0.06
	Q2 (61, 72]	18309	28.76	72.97 ± 0.11	57.29 ± 0.47	23.61 ± 0.30	33.44 ± 0.35	72.05 ± 0.08
	Q3 (72, 83]	18115	57.79	66.23 ± 0.13	66.87 ± 0.10	82.36 ± 0.21	73.81 ± 0.12	71.51 ± 0.05
	Q4 (83, 94]	18026	65.87	66.15 ± 0.14	67.84 ± 0.04	92.43 ± 0.35	78.25 ± 0.13	66.21 ± 0.13
Education Level	Q1 [0, 4]	18593	40.71	72.64 ± 0.08	65.91 ± 0.11	67.95 ± 0.07	66.91 ± 0.09	80.11 ± 0.03
	Q2 (4, 9]	18645	41.60	72.58 ± 0.09	66.56 ± 0.08	68.52 ± 0.24	67.53 ± 0.15	80.39 ± 0.04
	Q3 (9, 14]	18546	41.42	73.13 ± 0.18	67.19 ± 0.25	68.63 ± 0.13	67.90 ± 0.18	80.68 ± 0.05
	Q4 (14, 19]	18499	41.65	72.51 ± 0.07	66.59 ± 0.15	68.23 ± 0.15	67.40 ± 0.02	80.06 ± 0.06
FamilyHistory OfAlzheimer	No	52004	36.80	72.86 ± 0.05	62.22 ± 0.09	66.83 ± 0.10	64.44 ± 0.03	80.10 ± 0.05
	Yes	22279	51.95	72.39 ± 0.06	74.71 ± 0.08	70.82 ± 0.14	72.71 ± 0.08	78.47 ± 0.05
GeneticRiskFactor APOE-ε4 allele	No	59561	36.58	72.91 ± 0.05	62.66 ± 0.07	64.22 ± 0.15	63.43 ± 0.09	79.47 ± 0.04
	Yes	14722	60.61	71.93 ± 0.13	76.05 ± 0.12	78.37 ± 0.10	77.19 ± 0.10	78.23 ± 0.07
Hypertension	No	52134	41.38	72.59 ± 0.06	66.47 ± 0.10	68.14 ± 0.12	67.29 ± 0.06	80.16 ± 0.04
	Yes	22149	41.26	73.01 ± 0.05	66.80 ± 0.04	68.78 ± 0.12	67.77 ± 0.08	80.66 ± 0.04
Diabetes	No	59527	41.41	72.51 ± 0.04	66.35 ± 0.07	68.21 ± 0.09	67.26 ± 0.05	80.13 ± 0.04
	Yes	14756	41.08	73.55 ± 0.11	67.45 ± 0.19	68.85 ± 0.07	68.14 ± 0.09	81.04 ± 0.03

The subgroup results show generally similar performance across Gender, EducationLevel, Hypertension and Diabetes groups. The largest variations were observed across age groups, particularly in recall, which increased from 4.70% in the youngest group to 92.43% in the oldest group. Differences were also observed according to family history and APOE-ε4 allele status, indicating that model performance varies across some patient subgroups.

5.7. Misclassification and error analysis

To further investigate the moderate predictive performance, an error analysis was performed for XGBoost under the DWB configuration, focusing on false-positive (FP) and false-negative (FN) predictions. Figure 7 presents the predicted-probability distributions of the two true classes relative to the classification threshold of 0.50.

The probability distributions show substantial overlap between the two classes, particularly around the classification threshold, which helps explain the occurrence of both false-positive and false-negative predictions. To further characterize these errors, Table 15 reports the FPR and False Negative Rate (FNR) and distinguishes between borderline and high-confidence misclassifications. For descriptive error analysis, borderline FNs were defined by predicted probabilities of $0.40 \leq p < 0.50$ and borderline FPs by $0.50 \leq p < 0.60$, whereas high-confidence FNs and FPs were defined by $p < 0.30$ and $p \geq 0.70$, respectively. These probability ranges were used only for descriptive error analysis. Consistent with the overlap observed in Figure 7, the quantitative error

Table 15. Error analysis of XGBoost under the DWB configuration

Model	FPR (mean ± std, %)	FNR (mean ± std, %)	Borderline FP (mean ± std, %)	Borderline FN (mean ± std, %)	High-confidence FP (mean ± std, %)	High-confidence FN (mean ± std, %)
XGBoost+DWB	24.20 ± 0.09	31.67 ± 0.08	52.95 ± 0.29	33.71 ± 0.30	5.48 ± 0.18	19.07 ± 0.17

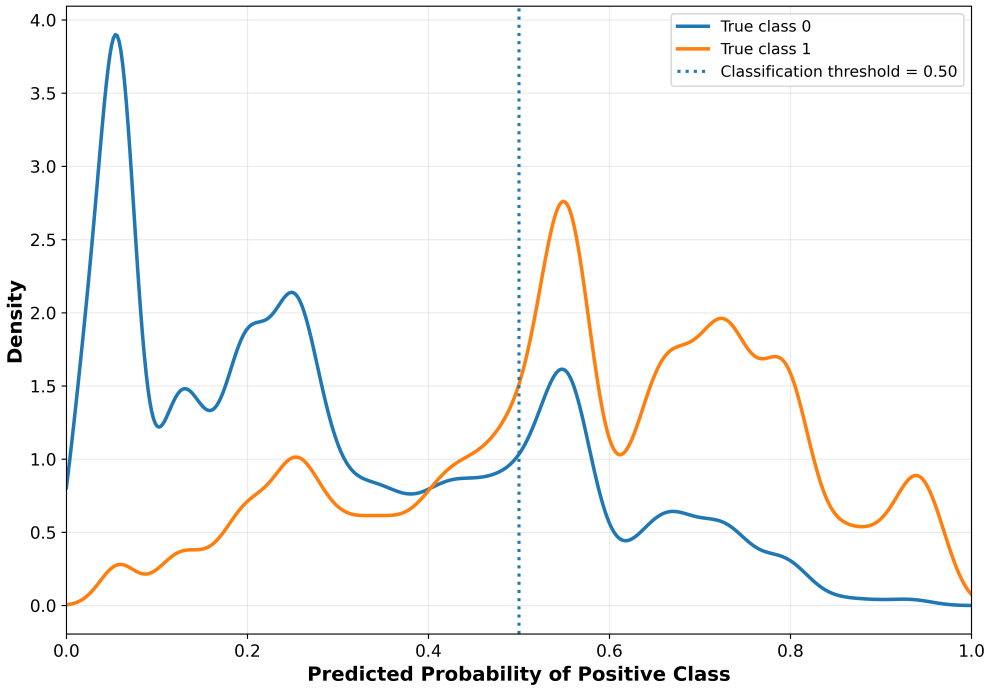


Figure 7. Predicted-probability distributions for XGBoost under the DWB configuration.

analysis showed that many misclassifications occurred near the classification threshold, particularly false positives, of which 52.95% were classified as borderline errors. The FNR (31.67%) was higher than the FPR (24.20%), while high-confidence false negatives (19.07%) were more frequent than high-confidence false positives (5.48%).

5.8. Component ablation analysis

To assess the contribution of the main pipeline components, an ablation analysis was performed using XGBoost, which achieved the highest AUC among the evaluated classifiers under DWB. Feature selection (FS), data balancing and hyperparameter tuning were evaluated individually and in combination using the same repeated nested cross-validation framework and identical outer validation partitions. The results are presented in Table 16. The ablation results show that feature selection and hyperparameter tuning both improved XGBoost performance compared with the baseline configuration. Their combination under DWB achieved the highest accuracy (72.72%) and AUC (80.31%), while also improving precision, recall and F1-score compared with the baseline. In contrast, adding data balancing increased recall and F1-score but reduced accuracy, precision and AUC. These results indicate that balancing mainly affected the precision–recall trade-off rather than consistently improving overall discrimination. Therefore, XGBoost with feature selection and hyperparameter tuning under DWB, hereafter referred to as FS-Tuned-XGBoost, was selected as the best-performing configuration based on its overall discrimination performance.

6. Discussion

The results obtained in this study, particularly the results of feature importance analysis, underscore the significant contribution of risk factors including lifestyle factors, age, education level and BMI to AD prediction. The study [3] reported that 12 modifiable risk factors can lead to AD, including limited education, hearing loss, hypertension,

Table 16. Component ablation analysis of XGBoost

Configuration	FS	Balancing	Tuning	Accuracy (mean ± std, %)	Precision (mean ± std, %)	Recall (mean ± std, %)	F1-score (mean ± std, %)	AUC (mean ± std, %)
Baseline	–	–	–	71.41 ± 0.13	65.72 ± 0.17	64.48 ± 0.19	65.10 ± 0.16	78.84 ± 0.06
FS only	Yes	–	–	72.14 ± 0.06	66.31 ± 0.11	66.29 ± 0.12	66.30 ± 0.06	79.61 ± 0.04
Balancing only	–	Yes	–	70.91 ± 0.05	61.76 ± 0.06	77.83 ± 0.11	68.87 ± 0.05	78.19 ± 0.09
Tuning only	–	–	Yes	72.17 ± 0.06	66.29 ± 0.07	66.53 ± 0.09	66.41 ± 0.07	79.86 ± 0.03
FS + Balancing	Yes	Yes	–	71.54 ± 0.04	62.98 ± 0.11	75.61 ± 0.43	68.72 ± 0.12	79.32 ± 0.09
FS + Tuning	Yes	–	Yes	72.72 ± 0.05	66.56 ± 0.08	68.33 ± 0.08	67.44 ± 0.05	80.31 ± 0.04
FS + Balancing + Tuning	Yes	Yes	Yes	71.74 ± 0.08	63.05 ± 0.15	76.42 ± 0.58	69.10 ± 0.17	79.55 ± 0.11

smoking, physical inactivity, depression, obesity, diabetes, loneliness, heavy drinking, head trauma, air pollution. However, several studies in the medical field have shown that the contributions of these risk factors differ and that some factors are more significant than others, which matches the findings of our feature-selection analysis. For instance, our study illustrates the low importance of smoking status feature, which aligns with the study [35] that concluded from eleven other investigations that there is no significant correlation between smoking and the risk of developing AD. Additionally, BMI and education level were identified as relevant features in our study and have a high association with the risk of developing AD, which is exhibited as reported in [36] and [3] respectively. Country was among the highly ranked features; however, it should not be interpreted as a causal risk factor for AD. The ablation analysis showed that its removal caused only a limited decrease in AUC while improving recall and F1-score. Given its non-clinical nature, Country may capture geographic, socioeconomic, healthcare-access, or data-collection differences rather than biological effects. Therefore, it should not be considered a clinically actionable predictor.

The present paper comes to complete and enrich our prior work on an optimized ML algorithm for AD diagnosis. The outcome of our previous work was a performant DT model thanks to tuning with grid search and application and validation under 5-fold cross-validation method [37]. In another recent work, we extended our research by integrating feature selection and explainable AI techniques and deploying the selected model in a web-based decision-support application for AD diagnosis [38]. Building on these previous studies, the present study adopts a broader and more systematic experimental framework, comparing multiple classifiers, feature-selection methods, and data-balancing techniques under repeated nested cross-validation.

Regarding the effect of data balancing, its impact on discrimination was limited and classifier-dependent. XGBoost achieved its highest AUC under DWB (80.31%), while the improvements observed for some other classifiers were small. This suggests that, for the moderate class imbalance in this dataset, balancing should be evaluated according to the classifier rather than applied systematically. Moreover, the statistical comparisons confirmed that XGBoost achieved a significantly higher AUC than the other evaluated classifiers under DWB, further supporting its selection as the reference classifier for subsequent analyses. A comparison of our best-performing model with previous studies using the same dataset or a similar tabular data modality has also been made (Table 17). Beyond this performance comparison, the main contribution of our approach lies in the systematic evaluation of feature selection, data balancing, hyperparameter optimization, and cross-validation within a common ML framework.

Compared to previous studies using the same dataset, our approach achieved the highest AUC (80.31%) and specificity (75.80%) among the studies reporting these metrics, while maintaining an accuracy of 72.72%. Although

Table 17. Comparison with recent studies using the same tabular AD dataset

Study	Dataset Type	Validation Strategy	Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)	AUC (%)
[13]	Tabular AD dataset [33]	Stratified 80/20 train–test split	DNN	73	72	76	–	74	–
[8]	Tabular AD dataset [33]	80/20 train–test split	GB	73	72	73	–	72	–
			LR	71	71	70	–	70	–
			RF	72	72	68	–	68	–
[14]	Tabular AD dataset [33]	80/20 train–test split; 10-fold stratified CV for tuning; final evaluation on held-out 20% test set	AdaBoost	72.97	65.37	73.63	72.50	69.25	80.00
Our study	Tabular AD dataset [33]	Repeated nested CV (5-fold outer CV × 5 repeats; 5-fold inner CV)	FS-Tuned-XGBoost	72.72	66.56	68.33	75.80	67.44	80.31

some studies reported higher accuracy or recall, differences in validation protocols should be considered when interpreting these results. Overall, our findings complement previous studies by providing a systematic evaluation of feature selection, data balancing, and hyperparameter optimization under a common validation framework. This framework also allows the contribution of each component to the final performance to be examined. The component ablation analysis showed that feature selection and hyperparameter tuning improved XGBoost performance, with their combination under DWB achieving the highest AUC (80.31%) and accuracy (72.72%). Adding data balancing increased recall and F1-score but reduced precision and AUC, indicating a trade-off rather than a consistent overall improvement. This trade-off is particularly relevant when considering the model for AD screening.

In this context, false negatives are particularly concerning because potential AD cases may be missed, which can delay further clinical assessment and follow-up. Therefore, high recall is important for reducing missed cases [39]. However, low precision can increase false positives, leading to unnecessary referrals and additional assessments. Thus, recall is prioritized in this screening context, while precision and specificity remain important for limiting unnecessary follow-up. These performance trade-offs should also be considered across different patient subgroups.

The subgroup analysis further revealed variations across age groups, particularly in recall. Differences across several other characteristics were smaller. These findings indicate that overall performance metrics do not fully reflect model behavior across patient subgroups.

The error analysis provided further insight into the moderate overall performance. It showed a higher false-negative rate (31.67%) than false-positive rate (24.20%), with high-confidence false negatives also more frequent than high-confidence false positives. Given the obtained accuracy and the remaining misclassification errors, the model is more appropriately considered as a potential support for AD screening and clinical decision-making rather than as a standalone diagnostic model.

Calibration and decision curve analyses complemented the performance evaluation by examining probability reliability and threshold-dependent net benefit. The calibration analysis showed that XGBoost provided the most favorable calibration, with the lowest Brier score and calibration parameters close to their ideal values. The DCA results further showed that most XGBoost configurations provided similar positive net benefit across a broad range of thresholds, while SMOTE-ENN showed reduced net benefit at higher thresholds. Overall, data balancing did not provide a consistent decision-analytic advantage over DWB.

Despite these contributions, several limitations should be acknowledged. First, the reliance on an imbalanced dataset may bias the learning process toward the predominant class, potentially reducing the detection performance for the underrepresented class. Second, external validation on an independent clinical dataset would provide further

evidence of the model's generalizability. However, comparable public AD tabular datasets containing similar features are currently limited and difficult to access.

7. Conclusion

Since no cure for AD currently exists, early and accurate diagnosis remains the most effective way to limit its progression. Our research contributes to this objective by developing a comprehensive ML framework for AD prediction. We integrated advanced ML methods KNN, LR, DT, RF, XGBoost and Bagging ME to predict AD. Experiments were performed using several data balancing techniques to evaluate their effect on predictive performance. XGBoost with feature selection and hyperparameter tuning under DWB provided the best overall discrimination, achieving 72.72% accuracy, 66.56% precision, 68.33% recall, 67.44% F1-score and 80.31% AUC. The results also showed that data balancing did not consistently improve discrimination and mainly affected the trade-off between precision and recall. Additionally, the present research identified several relevant variables in AD diagnosis and showed that their rankings varied across feature-selection methods and training partitions. The practical outcomes of our study may support clinical decision-making in the early identification of AD and stimulate additional research into developing high-performance models. Considering the obtained performance, the model may serve as a complementary tool for screening and clinical decision support. For future studies, higher performance could be achieved through the adoption of deep learning algorithms. The proposed approach could also be strengthened through clinical validation conducted in collaboration with medical professionals to ensure its practical relevance and reliability. Furthermore, the work could be extended to perform prediction using other kinds of data such as brain images, genes and biomarkers. Finally, external validation on independent datasets may be performed to further assess the robustness and generalizability of the developed models.

REFERENCES

1. S. Evans-Lacko et al., *World Alzheimer Report 2024: Global Changes in Attitudes to Dementia*, [Internet], 2024. Available from: https://www.alzint.org/resource/world-alzheimer-report-2024
2. World Health Organization, *World Health Organization official website-Demention and Alzheimer Key facts*, [Internet], 2025. Available from: https://www.who.int/news-room/fact-sheets/detail/dementia
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. S. El Khamlichi and L. Taidi, *Improving Machine Learning Performance Using Sampling Techniques for COVID-19 Imbalanced Data*, In *Intelligent Cybersecurity and Resilience for Critical Industries: Challenges and Applications*,1st ed., pp. 103-121, River Publishers, 2025. DOI: 10.1201/9788770047746-6.
5. S. El Khamlichi and L. Taidi, *Machine Learning Models for Predicting COVID-19 Mortality Using Epidemiological Features*, Statistics, Optimization and Information Computing, vol. 14, no. 2, pp. 677–703, 2025. DOI: 10.19139/soic-2310-5070-2159.
6. 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.
7. H. Alshamlan et al., *Identifying Effective Feature Selection Methods for Alzheimer's Disease Biomarker Gene Detection Using Machine Learning*, Diagnostics, vol. 13, no. 10, 1771, 2023. DOI: 10.3390/diagnostics13101771.
8. B. Singh et al., *Evaluating Machine Learning Algorithms for Early Detection of Alzheimer's Disease*, In Proceedings of 2025 Second International Conference on Cognitive Robotics and Intelligent Systems (ICC-ROBINS), pp. 510–515, 2025. DOI: 10.1109/ICC-ROBINS64345.2025.11086346.
9. 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.
10. J. De Velasco Oriol et al., *Benchmarking Machine Learning Models for Late-Onset Alzheimer's Disease Prediction from Genomic Data*, BMC Bioinformatics, vol. 20, no. 1, 709, 2019. DOI: 10.1186/s12859-019-3158-x.
11. G. Ahmed et al., *DAD-Net: Classification of Alzheimer's Disease Using ADASYN Oversampling Technique and Optimized Neural Network*, Molecules, vol. 27, no. 20, 7085, 2022. DOI: 10.3390/molecules27207085.
12. A. Zadgaonkar, R. Keskar, and O. Kakde, *Towards a Machine Learning Model for Detection of Dementia Using Lifestyle Parameters*, Applied Sciences, vol. 13, no. 19, 10630, 2023. DOI: 10.3390/app131910630.
13. S. A. Mohamed, M. A. A. E. M. Abu Resha, and M. G. Hassan, *Towards Transparent and Robust Early Diagnosis of Alzheimer's Disease: A Deep Learning and Visual Analytics Framework*, In Proceedings of 2025 Intelligent Methods, Systems, and Applications (IMSA), pp. 340–345, 2025. DOI: 10.1109/IMSA65733.2025.11167773.
14. H. Mahmoud, andF. Salama, *Leakage-Safe Boosting Ensembles With SHAP Explainability for Alzheimer's Disease Risk Prediction: A Rigorous Multi-Classifier Benchmark on Global Structured Clinical Data*, Sustainable Machine Intelligence Journal, vol. 14, no. 3, pp. 313–332, 2026. DOI: 10.63689/3005-3617.1092.

15. 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.
16. C. Lazar et al., *A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 4, pp. 1106–1119, 2012. DOI: 10.1109/TCBB.2012.33.
17. X. Cheng, *A Comprehensive Study of Feature Selection Techniques in Machine Learning Models*, Artificial Intelligence and Digital Technology, vol. 1, pp. 65–78, 2024. Available from: https://soapubs.com/index.php/AIDT/article/view/217
18. 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.
19. 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.
20. E. F. Agyemang et al., *Addressing Class Imbalance Problem in Health Data Classification: Practical Application From an Oversampling Viewpoint*, Applied Computational Intelligence and Soft Computing, vol. 2025, no. 1, 1013769, 2025. DOI: 10.1155/acis/1013769.
21. T. Elhassan et al., *Classification of Imbalance Data Using Tomek Link Combined with Random Under-Sampling as a Data Reduction Method*, Global Journal of Technology & Optimization, no. S1, 111, 2017. Available from: https://www.academia.edu/37371633/Classification_of_Imbalance_Data_using_Tomek_Link_T_Link_Combined_with_Random_Under_sampling_RUS_as_a_Data_Reduction_Method
22. Y. Han and I. Joe, *Enhancing Machine Learning Models Through PCA, SMOTE-ENN, and Stochastic Weighted Averaging*, Applied Sciences, vol. 14, no. 21, 9772, 2024. DOI: 10.3390/app14219772.
23. L. Chege, H. Waititu, and C. Nyakundi, *Comparative Analysis of Data Balancing Techniques in Mental Health Data: Application to Treatment Modalities*, International Journal of Statistics and Applications, vol. 14, no. 3, pp. 47–55, 2024. DOI: 10.5923/j.statistics.20241403.02.
24. M. Bansal, A. Goyal, and A. Choudhary, *A Comparative Analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory Algorithms in Machine Learning*, Decision Analytics Journal, vol. 3, 100071, 2022. DOI: 10.1016/j.dajour.2022.100071.
25. G. Ambrish et al., *Logistic Regression Technique for Prediction of Cardiovascular Disease*, Global Transitions Proceedings, vol. 3, no. 1, pp. 127–130, 2022. DOI: 10.1016/j.gltp.2022.04.008.
26. 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.
27. R. A. Azdy et al., *Performance Evaluation of Bagging Meta-Estimator in Lung Disease Detection: A Case Study on Imbalanced Dataset*, International Journal of Artificial Intelligence in Medical Issues, vol. 1, no. 2, 2023. DOI: 10.56705/ijaimi.v1i2.96.
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. J. C. Obi, *A Comparative Study of Several Classification Metrics and Their Performances on Data*, World Journal of Advanced Engineering Technology and Sciences, vol. 8, no. 1, pp. 308–314, 2023. DOI: 10.30574/wjaets.2023.8.1.0054.
30. V. B. Calster et al., *Calibration: the Achilles heel of predictive analytics*, BMC Medicine, vol. 17, no. 1, 230, 2019. DOI: 10.1186/s12916-019-1466-7.
31. L. Hoessly, *On misconceptions about the Brier score in binary prediction models*, Global Epidemiology, vol. 11, 100242, 2026. DOI: 10.1016/j.gloepi.2025.100242.
32. V. Vickers , and E. B. Elkin, *Decision Curve Analysis: A Novel Method for Evaluating Prediction Models*, Medical Decision Making, vol. 26, no. 6, pp. 565-574, 2006. DOI: 10.1177/0272989X06295361.
33. A. Panday, *Alzheimer's Prediction Dataset (Global)*, [Internet], 2025. Available from: https://www.kaggle.com/datasets/ankushpanday1/alzheimers-prediction-dataset-global
34. C. E. Fischer et al., *Determining the Impact of Psychosis on Rates of False-Positive and False-Negative Diagnosis in Alzheimer's Disease*, Alzheimer's & Dementia: Translational Research & Clinical Interventions, vol. 3, no. 3, pp. 385–392, 2017. DOI: 10.1016/j.trci.2017.06.001.
35. J. Li et al., *Risk Factors for Predicting Progression from Mild Cognitive Impairment to Alzheimer's Disease: A Systematic Review and Meta-Analysis of Cohort Studies*, Journal of Neurology, Neurosurgery & Psychiatry, vol. 87, no. 5, pp. 476–484, 2016. DOI: 10.1136/jnnp-2014-310095.
36. L. T. Nordestgaard, M. Christoffersen, and R. Frikke-Schmidt, *Shared Risk Factors between Dementia and Atherosclerotic Cardiovascular Disease*, International Journal of Molecular Sciences, vol. 23, no. 17, 9777, 2022. DOI: 10.3390/ijms23179777.
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 International Conference on Digital Age & Technological Advances for Sustainable Development (DATA 2025), pp. 182–192, 2026. DOI: 10.1007/978-3-032-22567-2_16.
38. A. Mettahri , S. El Khamlichi, M. El Haloui and B. Ettaki, *An Explainable Machine Learning–Based Web Tool for Alzheimer's Disease Diagnosis*, Statistics, Optimization & Information Computing, vol. 16, no. 3, pp. 1984–2000, 2026. DOI: 10.19139/soic-2310-5070-3624.
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.