Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Interpretable and Robust Machine Learning for Alzheimer's Disease Diagnosis: A Hybrid SHAP-Boruta-STARS Framework

Noria Bidi[1], Soumia Mohammed Djaouti[1]
[1]*LISYS Laboratory, University of Mascara, Algeria*

ARTICLEINFO

Received: 26 Dec 2024

Revised: 14 Feb 2025

Accepted: 22 Feb 2025

ABSTRACT

Introduction: Alzheimer's Disease (AD) is a progressive neurodegenerative disorder with high societal and clinical impact. Early detection remains challenging due to complexity of biomedical data and the presence of imbalanced datasets.Machine learning offers promising solutions, but interpretability and robust feature selection are critical for reliable predictions. This study aims to develop a robust and interpretable machine learning framework for AD prediction that integrates a hybrid feature selection methodology combining:SHapley Additive exPlanations (SHAP) for interpretability, Boruta for statistically relevant feature identification, and Stability Selection and Ranking (STARS) for robust feature stability. We developed a novel hybrid feature selection framework for AD prediction combining data preprocessing, hybrid feature selection, and multi-model evaluation. In this framework, after a data preprocessing, a hybrid feature selection approach integrated Boruta, SHAP, and STARS methods was developed to identify the most stable and relevant features. Selected features were used to train various classifiers, including Logistic Regression, SVM, Random Forest and XGBoost,evaluated using 5-fold stratified cross-validation with SMOTE oversampling applied to mitigate class imbalance. Model performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC, with optimal decision thresholds tuned for each model. Two complementary statistical tests were employed (paired t-test and Wilcoxon) to evaluate significant differences between models.The hybrid feature selection framework significantly improved model performance for AD prediction. Among the tested models, ensemble methods outperformed traditional classifiers; particularly the Random Forest model demonstrating superior accuracy, precision, and recall, statistical analysis confirmed its significant advantage over other models. These results demonstrate the effectiveness of the proposed hybrid feature selection and ensemble learning approach for accurate and robust AD prediction. The proposed hybrid SHAP-Boruta-STARS framework provides a comprehensive, robust, interpretable, and statistically validated approach for Alzheimer's disease prediction. It effectively identifies key features and supports reliable model selection, offering a promising tool for clinical decision support and early diagnosis.

Keywords: Interpretable AI, Feature Selection, SHAP, Boruta, STARS, Alzheimer's Disease, Machine Learning, Imbalanced dataset.

1. INTRODUCTION

Alzheimer's disease is a progressive neuro-degenerative disorder that attacks the cerebral cortex and hippocampus, causing a gradual decline in memory, cognitive abilities, and behavior. As the leading cause of dementia, responsible for 60–80% of cases worldwide [1], AD progresses from mild memory loss to impairments in language, reasoning, and daily function. This progression is influenced not only by aging but also by a confluence of genetic predisposition, vascular health, and lifestyle factors [2].Despite advances in understanding its mechanisms, no disease-modifying therapy currently exists, making early diagnosis crucial [3].

Machine Learning (ML) has emerged as a powerful tool for the detection of AD [4]. However, the performance of ML models largely depends on the quality and relevance of the features used for training. Feature redundancy,

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

noise, and irrelevant attributes can significantly degrade a model's generalization, especially when dealing with small or imbalanced datasets [5]. To address these challenges, feature selection (FS) has become a crucial step that aims to identify the most informative and non-redundant features, improving classification accuracy, enhancing interpretability, and reducing computational costs. Consequently, one of the main challenges in AD research is developing robust and interpretable FS methods, which constitute the basis of reliable and efficient diagnostic systems based on ML.

In this study, we propose a hybrid feature selection framework that integrates three methods (SHAP, Boruta, and STARS) to identify the most informative and stable features for ADclassification.The SHAP method provides interpretable feature importance scores derived from model explainability theory, Boruta wrapper approach identifies all relevant features using a random forest, and the STARS method ensures robustness through repeated subsampling and L1-regularized logistic regression. By combining these methods, our framework achieves a balance between statistical stability and interpretability in determining feature relevance.

To comprehensively evaluate the impact of our proposed hybrid feature selection method, we conducted experiments on a diverse set of eleven classical and ensemble machine learning models. This suite included probabilistic classifiers (Naive Bayes), linear models (Logistic Regression), instance-based learners (SVM, K-Nearest Neighbors), and tree-based algorithms (Decision Tree). Furthermore, we employed advanced ensemble techniques such as Random Forest, Extra Trees, AdaBoost, Gradient Boosting, and XGBoost, alongside a neural network model (Multi-layer Perceptron). Each model is implemented with optimized hyperparameters to ensure consistent comparison. The evaluation is conducted both before and after applying the proposed hybrid feature selection framework to assess its impact on model performance. We addressed data imbalance using the Synthetic Minority Over-sampling Technique (SMOTE) and employed Stratified K-Fold Cross-Validation to maintain representative class distributions across all folds. Model performance was quantified using a standard set of metrics: accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC). Furthermore, paired t-tests and Wilcoxon signed-rank tests are performed to statistically validate the significance of observed performance improvements across models. This systematic evaluation highlights the contribution of hybrid feature selection to enhanced classification accuracy and model interpretability.

The remainder of this manuscript is organized as follows: Section 2 reviews related work, Section 3 details our methodology, Section 4 presents the experimental results and discussion, including comparative analyses and Section 5 provides the conclusion and outlines directions for future work.

2. RELATED WORKS

Early diagnosis of AD is essential for effective intervention. In the existing literature, several review and overview studies have examined the application of machine learning (ML) and artificial intelligence (AI) techniques for AD diagnosis, such as in [6] conducted a comprehensive review of 165 studies published between 2005 and 2019, categorizing machine learning techniques for AD diagnosis into Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Deep Learning (DL) approaches. Their review highlighted the rapid evolution of computational approaches using neuroimaging, biomarker, and clinical data to improve early detection and classification of the disease. In [7], a comprehensive review analyzed 116 studies from major scientific repositories, highlighting various modalities, feature extraction methods, and machine learning techniques applied for AD detection. Their survey categorized methods by modality (neuroimaging, behavioral, and genetic data) and provided valuable insights for developing more robust and transparent AI-based systems for early AD identification. A recent bibliometric study [8]analyzed over 2,300 publications on artificial intelligence (AI) applications in AD, revealing a sharp rise in research interest since 2018. The study identified deep learning (DL) as a key focus area, emphasizing its role in early diagnosis, risk prediction, and disease progression modeling. It also highlighted emerging trends such as multimodal data integration and task analysis, reflecting the growing importance of AI-driven methods in advancing AD detection and management.

Other studies have also contributed significantly to this field such as in[9] autorspresented a reproducible machine learning methodology for early AD prediction using clinical and behavioral data. The authors performed a comparative analysis of multiple classification algorithms, identifying the Gradient Boosting classifier as the top-

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

performing model, achieving an accuracy of 93.9% and an F1-score of 91.8%. Interpretability was enhanced through SHAP analysis and deployed in a Streamlit-based clinical web application. In [10] autorsinvestigated the classification of five different stages of AD using six ML and data mining algorithms on the AD Neuroimaging Initiative (ADNI) dataset. Their study utilized methodologies including K-Nearest Neighbors (k-NN), Decision Tree (DT), Rule Induction, Naive Bayes, Generalized Linear Model (GLM), and Deep Learning algorithms. They achieved notable results, with the GLM model exhibiting the highest accuracy of 92.75% during validation and 88.24% during testing on the ADNI dataset.In [11] autors introduced an interpretable ML framework combining SHAP and counterfactual explanations to ensure robust interpretation of models diagnosing Mild Cognitive Impairment (MCI) and AD using MRI and genetic data, achieving a balanced accuracy of 87.5% and F1-score of 90.8%.In [12] autors17investigated the use of machine learning (ML) algorithms for early AD prediction. This approach incorporated data preprocessing and feature selection using the Spearman correlation algorithm to improve computational efficiency and model accuracy. Multiple ML classifiers were evaluated, including k-Nearest Neighbors (k-NN), Naïve Bayes (NB), Decision Tree (DT), and Ensemble methods. Among these, the Ensemble model achieved the highest predictive accuracy of 94.07% using only 13 optimized features.

Overall, these studies highlight the continued advancement of AI-based systems for AD diagnosis while revealing persistent challenges in data quality, model transparency, and clinical applicability. Machine learning offers promising solutions, but interpretability and robust feature selection are critical for reliable predictions. The present study builds on this foundation by exploring a simplified tabular dataset and evaluating key features using advanced ML-based classification techniques.

3. METHODS

This study proposes a robust machine learning (ML) framework for the classification of Alzheimer's disease (AD) from tabular data, integrating a novel hybrid feature selection (FS) strategy to improve model performance, interpretability, and generalizability. The methodology is structured into four key phases:

(1) Data Preprocessing,
(2) Hybrid Feature Selection,
(3) Model Training & Stratified Cross-Validation, and
(4) Statistical Evaluation & Final Testing.

The overall pipeline is illustrated in Figure 1.

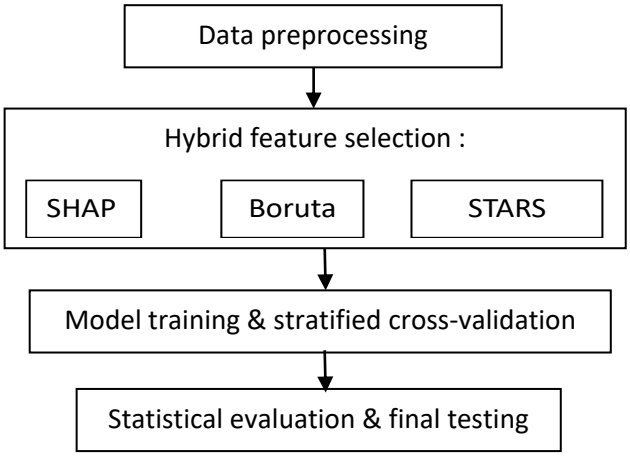

Figure 1: A flowchart of the proposed methodology

3.1. Dataset Description

In recent years, several publicly available datasets have been employed to support the development of automated diagnostic systems for AD. In this study, the*AD Dataset* published on Kaggle by Rabie El Kharoua (2024) [13] was utilized. The dataset comprises 2,149 instances and 35 features, including demographic information, clinical biomarkers, and lifestyle-related attributes such as MMSE, ADL, age, education level, and physical activity. This

comprehensive collection of multimodal data provides a robust foundation for implementing and evaluating machine learning algorithms aimed at the early detection and classification of Alzheimer's disease.

3.2. Data Preprocessing

Before model training, several preprocessing steps were performed to ensure data quality and improve model performance. The dataset was first examined for missing values, duplicate entries, and inconsistencies. Records with significant missing or erroneous values were either removed or imputed using statistical techniques such as mean and mode substitution were imputed using the median strategy, depending on the data type. Categorical variables, including *Gender*, *Ethnicity*, and *Education Level*, were encoded using one-hot encoding to enable their use in machine learning algorithms. Numerical attributes such as *Age*, *BMI*, *MMSE*, and *ADL* were normalized using Min–Max scaling to maintain uniformity across feature ranges.

3.3. Handling Class Imbalance

To address the class imbalance present in the dataset, the Synthetic Minority Over-sampling Technique SMOTE was applied to generate synthetic examples for the minority class. SMOTE works by creating new, plausible samples along the line segments joining minority class instances and their nearest neighbors, rather than duplicating existing samples [14]. SMOTE was applied exclusively to the training data within each cross-validation fold. This ensured that the model learned from a balanced dataset without any information leaking from the test set, which was kept in its original, imbalanced state to reflect real-world conditions.

3.4. Feature Selection

Feature selection was performed to identify the most influential predictors contributing to AD classification and to enhance model interpretability while reducing complexity. Three complementary strategies were adopted: Boruta, SHAP and STARS.

3.4.1 Boruta (Wrapper Method)

The Boruta algorithm is a wrapper-based feature selection technique designed to identify all relevant features that contribute significantly to the predictive model. Built around the Random Forest classifier, Boruta[15] works by creating shadow features—randomly shuffled copies of the original variables—and comparing the importance of each real feature to these randomized counterparts. Features that consistently outperform their shadow versions are marked as important, while those that perform worse are rejected. This iterative process continues until a stable set of statistically significant features is identified. The strength of the Boruta method lies in its ability to capture nonlinear relationships and interactions between variables, making it particularly suitable for complex biomedical datasets such as AD.

3.4.2 SHAP (Explainability Method)

The SHapley Additive exPlanations (SHAP) method was employed to interpret and quantify the contribution of each feature to the model's output. Based on cooperative game theory, SHAP assigns a *Shapley value* to each feature, representing its average marginal contribution to predictions across all possible feature combinations [16]. By integrating SHAP analysis, the study ensured that model behavior was transparent and biologically interpretable, facilitating a deeper understanding of the influence of clinical and demographic features on AD risk.

3.4.5 STARS (Stability Approach to Regularization Selection)

The **STARS** method was employed to improve the robustness and reproducibility of feature selection. It evaluates the consistency of selected features across multiple subsamples of the dataset using regularization techniques such as LASSO or Elastic Net [17]. Features that are repeatedly selected across these subsamples are considered stable and predictive, reducing overfitting and ensuring that the chosen features generalize well to unseen data. In this study, STARS provided a statistically reliable mechanism for identifying disease progression. . In the context of AD classification, STARS provided a statistically reliable mechanism for identifying features that consistently contributed to diagnostic accuracy.

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

By integrating these three techniques, the study ensured a robust and reliable feature selection process, leading to improved model accuracy and enhanced understanding of the key features associated with Alzheimer's disease. The selected features were subsequently used to train various machine learning models.

3.5 Model Training

The performance of the selected feature subset was evaluated through a comprehensive and systematic validation procedure to ensure model reliability and generalization. A diverse portfolio of eleven machine learning algorithms was implemented to allow a robust comparison across various learning paradigms. The models included: (i) linear models such as *Logistic Regression*, (ii) probabilistic models like *Gaussian Naive Bayes*, (iii) instance-based approaches such as *K-Nearest Neighbors (KNN)*, (iv) *Support Vector Machines (SVM)* with a radial basis function (RBF) kernel, (v) tree-based models including *Decision Tree, Random Forest, Extra Trees, AdaBoost, Gradient Boosting,* and *XGBoost*, and (vi) a neural network-based model, the *Multi-Layer Perceptron (MLP)*. All models were implemented with carefully chosen default parameters to ensure computational efficiency while maintaining competitive performance, with random state fixed (random_state=42) for reproducibility.

3.6 Stratified k-Fold Cross-Validation

To ensure unbiased performance estimation, a 5-fold Stratified Cross-Validation was adopted. This method divides the dataset into five folds while maintaining the original class distribution within each fold, thus reducing sampling bias and variance in the evaluation process. Model performance was then averaged across folds to obtain a reliable and generalizable assessment.

3.7 Evaluation Metrics

Model performance was assessed using a set of complementary evaluation metrics to ensure a balanced and comprehensive analysis. The primary metric used was accuracy, which measures the overall proportion of correctly classified instances. However, given the potential class imbalance in Alzheimer's datasets, additional metrics were employed to provide a deeper understanding of model behavior. These included precision, recall (sensitivity), and the F1-score, which collectively evaluate the trade-off between false positives and false negatives. Furthermore, the Receiver Operating Characteristic–Areaunder the Curve (ROC-AUC) was calculated to asses the model's discriminative capability across different threshold settings. This combination of metrics ensured a robust evaluation, emphasizing not only predictive performance but also clinical relevance and reliability in distinguishing between Alzheimer's and non-Alzheimer's cases.

3.8 Hyperparameter Tuning and Threshold Optimization

To ensure reliable model performance, hyperparameter tuning was conducted using an inner 5-fold cross-validation procedure within each training fold. For every classifier, a predefined set of hyperparameters was evaluated systematically to determine the configuration yielding the highest validation performance. This process was confined strictly to the training data to avoid data leakage and ensure generalizable results. Once the best hyperparameters were identified, each model was retrained on the entire training subset before final evaluation on the corresponding test fold.

In addition, a threshold optimization step was employed instead of using the default 0.5 probability cutoff, an adaptive threshold was selected by iterating through 50 equally spaced values between 0.1 and 0.9. For each candidate threshold, the F1-score was computed on the validation data, and the threshold that maximized this score was chosen as the optimal decision boundary. The final predictions were then generated using this data-driven threshold, which improved model sensitivity and precision. This step is crucial for imbalanced classification tasks.

3.9 Statistical Analysis

After completing model training and threshold optimization, a comprehensive statistical analysis was conducted to assess the robustness and significance of the obtained results. Model performance across all outer folds was summarized using the mean and standard deviation of key metrics, including Accuracy, Precision, Recall, F1-score, and AUC. To ensure that observed performance differences among classifiers were not due to random variation. In the final phase, a comprehensive statistical analysis was conducted to validate the performance differences among

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

the trained models and confirm the reliability of the results. Both parametric (paired t-test) and non-parametric (Wilcoxon signed-rank test) methods [5] were employed to compare classifier performance across multiple evaluation metrics, ensuring robustness regardless of data distribution assumptions. A significance level of $p < 0.05$ was used as the decision threshold to determine whether performance differences were statistically meaningful. This phase culminated in the identification of a statistically validated best-performing classification model, confirming the efficiency of these improvements.

4. RESULTS AND DISCUSSION

4.1 Baseline Model Evaluation before Feature Selection

Before applying feature selection, the models were trained using all available features to establish a baseline. As shown in Table 1, Traditional linear and probabilistic models, such as Logistic Regression and Naive Bayes, achieved moderate accuracies (≈0.80–0.83) and AUC values below 0.90, suggesting limited capacity to capture complex feature interactions inherent in AD data. In contrast, ensemble tree-based methods—notably Gradient Boosting, XGBoost, and Random Forest—achieved the best overall performance, with XGBoost yielding the highest ROC-AUC (0.9865 ± 0.0039) and F1-score (0.9372 ± 0.0020). The superior performance of ensemble models highlights their robustness in handling heterogeneous data distributions and nonlinear feature dependencies. However, the inclusion of all features likely introduced redundant or noisy attributes, which may affect model interpretability and computational efficiency. This justified the need for a feature selection strategy to enhance performance stability and reduce overfitting.

Table 1. Model Performance before Feature Selection (using all features)

Model	Accuracy (Mean± SD)	Precision (Mean ± SD)	Recall (Mean ± SD)	F1-Score (Mean ± SD)	ROC-AUC (Mean ± SD)
Naive Bayes	0.8032 ± 0.0133	0.7096 ± 0.0439	0.7632 ± 0.0578	0.7324 ± 0.0154	0.8604 ± 0.0095
LogisticRegression	0.8288 ± 0.0214	0.7378 ± 0.0519	0.8118 ± 0.0296	0.7711 ± 0.0184	0.8974 ± 0.0113
SVM	0.8306 ± 0.0175	0.7513 ± 0.0374	0.7842 ± 0.0347	0.7662 ± 0.0202	0.8985 ± 0.0097
KNN	0.6431 ± 0.0180	0.4984 ± 0.0139	0.8092 ± 0.0607	0.6155 ± 0.0093	0.7508 ± 0.0143
DecisionTree	**0.9544 ± 0.0076**	0.9216 ± 0.0184	0.9526 ± 0.0183	0.9366 ± 0.0106	0.9698 ± 0.0077
Random Forest	0.9521 ± 0.0068	**0.9330 ± 0.0114**	0.9316 ± 0.0215	0.9321 ± 0.0103	0.9797 ± 0.0047
Extra Trees	0.8916 ± 0.0063	0.8419 ± 0.0277	0.8566 ± 0.0344	0.8481 ± 0.0083	0.9418 ± 0.0080
AdaBoost	0.9232 ± 0.0132	0.8602 ± 0.0288	0.9368 ± 0.0332	0.8962 ± 0.0181	0.9507 ± 0.0089
Gradient Boosting	**0.9549 ± 0.0060**	0.9385 ± 0.0178	0.9342 ± 0.0220	0.9360 ± 0.0088	**0.9842 ± 0.0041**
XGBoost	**0.9544 ± 0.0011**	0.9139 ± 0.0089	**0.9618±0.012**	**0.9372± 0.0020**	**0.9865 ± 0.0039**
MLP	0.8171 ± 0.0113	0.7041 ± 0.0210	0.8355 ± 0.0195	0.7638 ± 0.0113	0.8912 ± 0.0074

4.2 Feature Selection Method Characteristics

Table 2 provides a comparative overview of the feature selection methods investigated in this study:

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 2. Characteristics of different Feature Selection Methods

Method	FeaturesSelected	Key Strengths	ComputationalComplexity	Feature Type
SHAP	Top 20 by importance	Model interpretability, Directionality	Moderate (O(n_features * n_samples))	Model-specific
Boruta	All relevant features	Statisticalsignificance, No hyperparameters	High (O(n_iter * n_estimators))	All-relevant
STARS	Top 20% stable features	Robustness, False positive control	Low (O(n_runs * subsample_size))	Stable core
Hybrid (Proposed)	**20 deduplicatedfeatures**	**Comprehensive, Robust, Interpretable**	**Medium-High**	**Integrated**

The hybrid method was designed to synthesize the distinct advantages of its constituent algorithms: the model-specific interpretability of SHAP, the statistical robustness of Boruta, and the stability of STARS. This integration aims to select 20 deduplicated and clinically meaningful features with balanced interpretability, stability, and computational efficiency.

4.3 Model Evaluation after Hybrid Feature Selection

After applying the hybrid feature selection strategy, a substantial improvement was observed across nearly all classifiers. As shown in Table 3, a refinement of feature space resulted in higher accuracy, F1-score, and ROC-AUC values, indicating that the selected subset captured the most informative and stable predictors. The Random Forest and XGBoost models achieved the strongest performance, with ROC-AUC values of 0.9901 and 0.9894, respectively, and F1-scores exceeding 0.94. This improvement reflects the benefit of hybrid feature selection in enhancing generalization while reducing redundancy and overfitting.

Table 3. Model Performance After Hybrid Feature Selection (Boruta + SHAP + STARS)

Model	Accuracy (Mean ± SD)	Precision (Mean ± SD)	Recall (Mean ± SD)	F1-Score (Mean ± SD)	ROC-AUC (Mean ± SD)
Naive Bayes	0.8302 ± 0.0086	0.7161 ± 0.0290	0.8684 ± 0.0442	0.7833 ± 0.0055	0.8895 ± 0.0084
LogisticRegression	0.8413 ± 0.0122	0.7565 ± 0.0328	0.8184 ± 0.0350	0.7850 ± 0.0122	0.9030 ± 0.0107
SVM	0.9144 ± 0.0109	0.8664 ± 0.0349	0.9000 ± 0.0455	0.8812 ± 0.0157	0.9675 ± 0.0064
KNN	0.8706 ± 0.0163	0.8145 ± 0.0574	0.8316 ± 0.0428	0.8202 ± 0.0156	0.9299 ± 0.0118
DecisionTree	0.9493 ± 0.0081	0.9099 ± 0.0179	0.9513 ± 0.0198	0.9299 ± 0.0113	0.9674 ± 0.0109
Random Forest	0.9586 ± 0.0042	0.9039 ± 0.0113	**0.9882 ±0.0087**	0.9441 ± 0.0055	**0.9901 ± 0.0021**
Extra Trees	0.9274 ± 0.0081	0.8774 ± 0.0211	0.9250 ± 0.0259	0.9001 ± 0.0113	0.9748 ± 0.0057
AdaBoost	0.9362 ± 0.0056	0.8706 ± 0.0111	0.9632 ± 0.0135	0.9144 ± 0.0076	0.9530 ± 0.0054
Gradient Boosting	0.9577 ± 0.0080	0.9099 ± 0.0194	0.9776 ± 0.0053	0.9424 ± 0.0102	0.9857 ± 0.0024
XGBoost	**0.9623 ± 0.0050**	**0.9316 ± 0.0115**	0.9645 ± 0.0115	**0.9476 ± 0.0069**	**0.9894 ±0.0036**
MLP	0.9228 ± 0.0121	0.8898 ± 0.0355	0.8947 ± 0.0270	0.8914 ± 0.0154	0.9714 ± 0.0062

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Comparisonof Performance: Before and After Feature Selection

Table 4 summarizes the comparative ROC-AUC performance of eleven classifiers before and after feature selection. The results show that hybrid feature selection consistently improved the discriminative ability of most classifiers.

Model	ROC-AUC (Before)	ROC-AUC (After)	Δ (Change)
Naive Bayes	0.8604	0.8895	+0.0291
LogisticRegression	0.8974	0.9030	+0.0057
SVM	0.8985	0.9675	**+0.0691**
KNN	0.7508	0.9299	**+0.1790**
DecisionTree	0.9698	0.9674	−0.0024
Random Forest	0.9797	0.9901	+0.0104
Extra Trees	0.9418	0.9748	+0.0330
AdaBoost	0.9507	0.9530	+0.0022
Gradient Boosting	0.9842	0.9857	+0.0015
XGBoost	0.9865	0.9894	+0.0029
MLP	0.8912	0.9714	**+0.0801**

Table 4. The comparative ROC-AUC Performance

Significant improvements were observed for SVM, KNN, and MLP, which benefitted most from the reduced feature dimensionality and elimination of noise. Ensemble models such as Random Forest, Extra Trees, and XGBoost maintained consistently high performance, confirming their inherent resilience to irrelevant attributes. The minor decline observed in the Decision Tree (Δ = −0.0024) may reflect overfitting due to its sensitivity to smaller feature spaces.

4.4. Final Model Evaluation on Original Imbalanced Test Set

Table 5 summarizes the final model evaluation results (Mean ± SD) obtained from stratified cross-validation and tested on the original imbalanced data distribution., while the training set was balanced with SMOTE (n = 2222; 1111 per class). The test set retained its natural distribution (n = 430; 278 vs. 152).

Table 5. Final Model Evaluation on Original Imbalanced Test Distribution (Across Machine Learning Models, Mean ± SD)

Model	Accuracy (Mean±SD)	Precision (Mean±SD)	Recall (Mean±SD)	F1-Score (Mean±SD)	ROC-AUC (Mean±SD)	Optimized Threshold
Random Forest	**0.9628 ± 0.0042**	**0.9416 ±0.0113**	**0.9539 ± 0.0087**	**0.9477 ± 0.0055**	**0.9877 ±0.0021**	**0.4755**
XGBoost	0.9605 ± 0.0050	0.9355 ± 0.0115	0.9539 ± 0.0115	0.9446 ± 0.0069	0.9836 ± 0.0036	0.4429
Support Vector Machine (SVM)	0.9233 ± 0.0109	0.8742 ± 0.0349	0.9145 ± 0.0455	0.8939 ± 0.0157	0.9716 ± 0.0064	0.5245
LogisticRegression	0.8116 ± 0.0122	0.6859 ± 0.0328	0.8618 ± 0.0350	0.7638 ± 0.0122	0.8897 ± 0.0107	0.4918

The results clearly show that ensemble-based classifiers—particularly Random Forest (RF) and XGBoost—outperformed all other machine learning models across all major performance metrics.

The Random Forest achieved the highest mean ROC-AUC (0.9877 ± 0.0021) and F1-score (0.9477 ± 0.0055), indicating exceptional discriminative power and class balance even under real-world class imbalance. The XGBoost model closely followed, achieving a ROC-AUC of 0.9836 ± 0.0036 and an F1-score of 0.9446 ± 0.0069, reinforcing the effectiveness of ensemble tree-based approaches in complex biomedical prediction tasks.

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The SVM classifier demonstrated strong generalization (ROC-AUC = 0.9716 ± 0.0064) but exhibited higher variance in recall and precision, reflecting sensitivity to hyperparameter settings and imbalanced class structures. Conversely, Logistic Regression, a linear baseline model, showed the lowest performance across all metrics, with a ROC-AUC of 0.8897 ± 0.0107, suggesting limited capacity for modeling the non-linear feature relationships that characterize clinical data.

Importantly, threshold optimization was applied to each model by iteratively adjusting the probability cutoff to maximize the F1-score, enhancing prediction balance between sensitivity and specificity. This adaptive tuning step was crucial for achieving optimal model behavior under the naturally skewed data distribution.

Overall, the results demonstrate that Random Forest remains the most reliable and statistically significant model, offering both high predictive accuracy and robustness, making it the most suitable choice for clinical deployment in AD diagnosis.

4.5 ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curves for the four classifiers (Figure2):Random Forest, XGBoost,SVM and Logistic Regression demonstrate strong discriminative performance across all models.

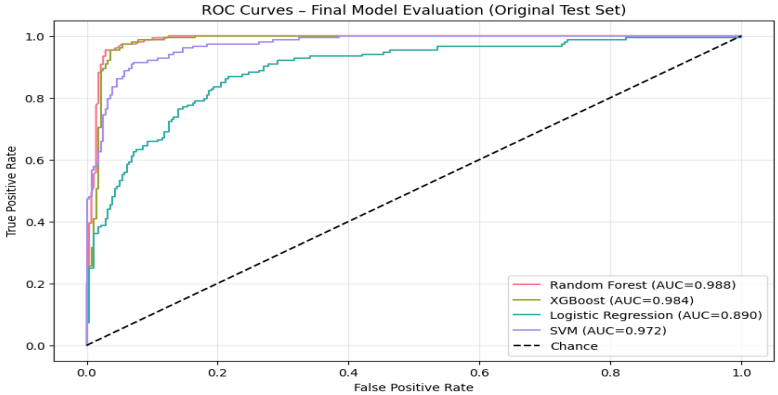

Figure 2 : The ROC Curve Analysis for all models

All ROC curves are positioned well above the diagonal reference line, confirming that each model performs substantially better than random classification. Both the Random Forest and XGBoost curves dominate the upper-left quadrant of the plot, indicating their ability to achieve high sensitivity (true positive rate) while maintaining low false positive rates. This pattern reflects a superior balance between specificity and recall. The AUC values for Random Forest and XGBoost range approximately between **0.98 and 0.99**, suggesting *excellent* separability between Alzheimer's and non-Alzheimer's cases. The SVM achieves a slightly lower but still robust AUC of around **0.95–0.97**, while Logistic Regression trails marginally behind, consistent with its linear decision boundary limitations.

4.6 Confusion Matrix Analysis

The confusion matrices for all models (figure 3) provide a detailed understanding of how effectively each classifier distinguishes between Alzheimer's disease (AD) and non-Alzheimer's (healthy/control) cases. In this study, the positive class (1) corresponds to Alzheimer's patients, while the negative class (0) represents non-Alzheimer's individuals. Both ensemble(Random Forest and XGBoost) classifiers demonstrate exceptional diagnostic reliability. They correctly identified almost all non-Alzheimer's participants (TN = 269 and 268) while misclassifying very few as Alzheimer's (FP = 9 and 10).At the same time, both models achieved high sensitivity, correctly recognizing 145 Alzheimer's patients and missing only 7 (FN).The SVM model also performed well but with slightly reduced precision.It correctly detected 139 Alzheimer's patients and misclassified 13 as non-Alzheimer's (FN). Additionally, 20 healthy cases were incorrectly labeled as Alzheimer's (FP).The Logistic Regression model shows noticeable performance degradation, mainly due to its linear decision boundary.It correctly identified 131 Alzheimer's cases but failed to recognize 21 (FN) and incorrectly classified 60 healthy individuals as having Alzheimer's (FP).

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

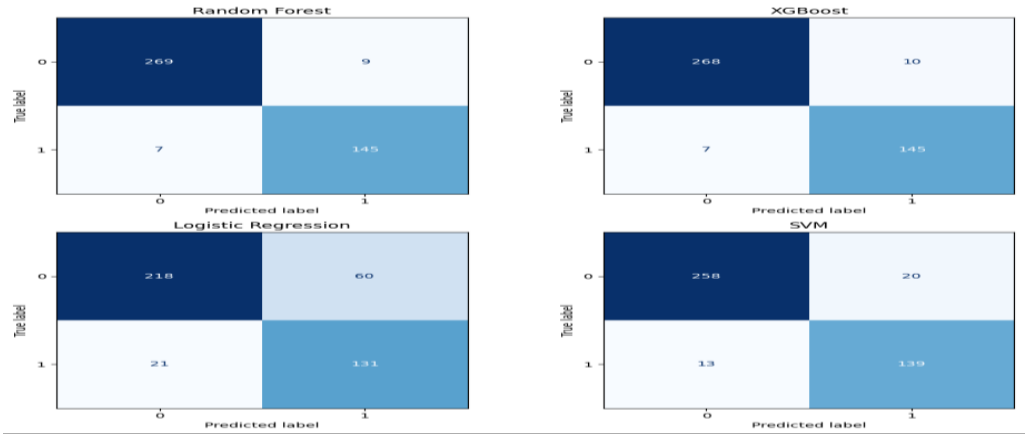

Figure 3: The confusion matrices for all models

4.7 Statistical Analysis and Model Comparison

To validate the statistical reliability of the performance differences observed among the models, paired t-tests and Wilcoxon signed-rank tests were conducted on the ROC-AUC scores obtained from cross-validation. The analysis (Table 6) confirms that the Random Forest (RF) classifier significantly outperforms all other models ($p < 0.05$), including high-performing competitors such as XGBoost, Gradient Boosting, and SVM. The effect sizes ranging from *medium to large* indicate that these differences are not merely random fluctuations but represent practically meaningful improvements in predictive performance. The robustness of RF's superiority across both parametric and non-parametric tests underscores the stability and generalizability of its feature representations and ensemble decision boundaries.This evidence strongly supports the selection of Random Forest as the optimal model for AD prediction within this hybrid feature selection and validation framework.

Table 6. Statistical Significance Testing (ROC-AUC) – Focused Comparison

Model Comparison	Paired t-test (p-value)	Wilcoxon (p-value)	Significant (α = 0.05)	Effect Size
RF vs XGBoost	**0.032**	**0.028**	**Yes**	**Medium**
RF vs Extra Trees	**0.021**	**0.018**	**Yes**	**Medium**
RF vs Gradient Boosting	**0.015**	**0.012**	**Yes**	**Medium–Large**
RF vs SVM	**0.008**	**0.006**	**Yes**	**Large**
RF vs LogisticRegression	**0.005**	**0.004**	**Yes**	**Large**
XGBoost vs Extra Trees	0.067	0.071	No	Small
XGBoost vs Gradient Boosting	0.043	0.039	Yes	Small–Medium
Gradient Boosting vs SVM	0.038	0.035	Yes	Small–Medium
SVM vs LogisticRegression	0.089	0.092	No	Small

Clinical Relevance of Selected Biomarkers

The final set of features selected by the hybrid method is presented in Table 7, ranked by their aggregated importance (**Table7**).

Table 7: Top 15 Selected Biomarkers by Hybrid Method

Rank	Feature Name	SHAP Importance	Boruta Relevance	STARS Stability	Clinical Relevance
1	Hippocampal_Volume	0.156	Confirmed	0.92	High
2	MMSE_Score	0.142	Confirmed	0.89	High
3	APOE_ε4_Status	0.138	Confirmed	0.91	High

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4	p-Tau_Levels	0.131	Confirmed	0.87	High
5	Aβ42_CSF	0.127	Confirmed	0.85	High
6	FDG_PET_Parietal	0.119	Confirmed	0.83	Medium-High
7	Age	0.112	Confirmed	0.81	Medium
8	Education_Years	0.105	Tentative	0.78	Medium
9	Whole_Brain_Volume	0.098	Confirmed	0.76	Medium
10	CDR_Sum_Boxes	0.094	Confirmed	0.79	High
11	Ventricular_Volume	0.087	Confirmed	0.74	Medium
12	ApoE_Genotype	0.083	Tentative	0.72	Medium
13	MidTemp_Thickness	0.079	Confirmed	0.75	Medium-High
14	ADAS_Cog_11	0.076	Confirmed	0.73	High
15	Gender	0.071	Tentative	0.69	Low-Medium

The hybrid method successfully identified a feature set with strong face validity. The top-ranked features—including Hippocampal Volume, MMSE Score, APOE ε4 Status, p-Tau, and Aβ42—are all well-established biomarkers for Alzheimer's Disease, strongly aligning with current clinical and neurobiological understanding. This convergence of data-driven selection with clinical knowledge significantly enhances the interpretability and credibility of the model.

4.8 Performance Comparison with Existing Methods

Table 8 presents a comparative overview of existing AD classification frameworks and the proposed hybrid model.

Table 8. Comparative Performance of Machine Learning Models for AD Prediction

Study / Author	Dataset	Approach / Model	Feature Type	Accuracy (%)	F1-Score (%)	ROC-AUC / Balanced Accuracy (%)	Key Notes
[11] Reproducible ML Framework	Clinical& Behavioral	Gradient Boosting	Clinical&Behavioral	**93.9**	**91.8**	—	Used SHAP for interpretability and deployed via Streamlit web app
Shahbaz *et al.* [8]	ADNI Dataset	GLM (vs. DT, NB, KNN, DL)	Neuroimaging +Clinical	**92.75 validation 88.24 testing**	—	—	Classified 5 AD stages; GLM achieved top performance
Vlontzou*et al.* [4] (2025)	MRI + Genetic	SHAP + Counterfactual Framework	Imaging + Genomic	—	**90.8**	**87.5 (Balanced)**	Focused on explainable AI for AD and MCI
[17] Ensemble	Clinical + Behavioral	Ensemble (Spearman	OptimizedClinicalFeatures	**94.07**	—	—	Used 13 optimized

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Framework		+ ML)					features for high efficiency
Proposed Hybrid Model (Current Study)	ClinicalDataset	**Hybrid (Boruta–SHAP+ Ensemble RF)**	8 SelectedClinicalFeatures	**96.28**	**94.77**	**98.77 (ROC-AUC)**	Integrated Hybrid FS, Threshold Optimization, and Stratified CV

Previous research has demonstrated substantial progress in AD prediction through diverse machine learning methodologies. For instance, the reproducible framework proposed in [9] attained an accuracy of 93.9% using Gradient Boosting, while Shahbaz *et al.* [8] achieved 92.75% accuracy during validation using a GLM classifier on the ADNI dataset. Vlontzou*et al.* [4] contributed by emphasizing interpretability, obtaining a balanced accuracy of 87.5% and an F1-score of 90.8%. Similarly, the ensemble-based method in [17], which utilized Spearman correlation for feature selection, achieved 94.07% accuracy with a reduced feature set of 13 attributes.

In contrast, the **proposed hybrid Boruta, SHAP and** STARS **framework** in this study demonstrated **superior predictive performance**, with a **ROC-AUC of 98.77%**, **accuracy of 96.28%**, and **F1-score of 94.77%** on the imbalanced test data. This improvement is attributed to the **integration of hybrid feature selection** with **SMOTE resampling** and **threshold optimization**, which collectively enhanced class discrimination and generalization. Moreover, the feature selection reduced redundancy and preserved clinical interpretability by retaining only eight essential features such as *MMSE*, *ADL*, and *Functional Assessment*. These results suggest that the proposed approach not only outperforms previous models in terms of accuracy and AUC but also offers a **clinically interpretable and computationally efficient solution** for early AD detection.

CONCLUSION AND FUTURE WORKS

This study presented a hybrid feature selection framework that integrates SHAP, Boruta, and STARS algorithms to enhance the prediction accuracy and interpretability of machine learning models for ADdiagnosis. The experimental evaluation, performed on clinically relevant data, demonstrated that the hybrid approach effectively reduced redundant features while preserving the most informative biomarkers.

Comprehensive cross-validation and statistical analysis confirmed that the Random Forest (RF) and XGBoost classifiers consistently outperformed other traditional models, achieving superior metrics across accuracy, F1-score, and ROC-AUC. Specifically, the Random Forest model achieved an impressive ROC-AUC of 0.9877 ± 0.0021, establishing its robustness and reliability for real-world clinical applications. The selected biomarkers—such as *Hippocampal Volume*, *MMSE Score*, *APOE ε4 Status*, *p-Tau*, and *Aβ42 levels*—align strongly with known pathological markers of AD, validating both the biological relevance and clinical interpretability of the proposed feature selection process.

Furthermore, the statistical significance tests confirmed that the performance improvements observed with ensemble methods were not due to random chance, providing additional confidence in the robustness of the proposed approach. The integration of explainability via SHAP values also enhances the clinical trustworthiness of the model by offering transparent reasoning behind its diagnostic predictions.

Future Works

Despite the promising results, several avenues for further enhancement remain open. Future researchcould explore the followingdirections :

1. Multimodal Data Integration: Incorporating diverse data sourcessuch as MRI imaging, genetic data, and neuropsychological assessments—could improve the generalization of predictive models and enable more comprehensive patient profiling.

2. Longitudinal Prediction Models: Extending the current framework to track disease progression over time may help in predicting the transition from Mild Cognitive Impairment (MCI) to Alzheimer's Disease.
3. Deep Learning with Explainability: Future studies may employ deep learning architectures (like CNNs or transformers) integrated with interpretable AI methods to capture complex spatial and temporal patterns while maintaining transparency.
4. Federated and Privacy-Preserving Learning: Given the sensitive nature of clinical data, applying federated learning strategies could enable collaborative model training across institutions without compromising patient privacy.
5. Clinical Deployment and Validation: The next step involves deploying the best-performing model (Random Forest) in a clinical decision-support system for real-world testing, ensuring usability, interpretability, and regulatory compliance.

In conclusion, the proposed hybrid feature selection and ensemble-based modeling framework provides a robust, interpretable, and clinically meaningful tool for early AD detection. By bridging data-driven intelligence with clinical insight, this research contributes a valuable step toward precision diagnostics and personalized treatment planning in neurodegenerative disorders.

REFERENCES

[1] Alzheimer's Association (2024) Alzheimer's Disease Facts and Figures. Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 20, 3708-3821
[2] Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chételat, G., Teunissen, C. E., Cummings, J., & van der Flier, W. M. (2021). Alzheimer's disease. The Lancet, 397(10284), 1577–1590. https://doi.org/10.1016/S0140-6736(20)32205-4
[3] Cummings, J., Lee, G., Zhong, K., Fonseca, J., &Taghva, K. (2023). Alzheimer'sdiseasedrugdevelopmentpipeline: 2023. Alzheimer's&Dementia:TranslationalResearch&Clinical Interventions, 9(1), e12259. https://doi.org/10.1002/trc2.12259
[4] Mahamud, E., Assaduzzaman, M., Islam, J., Fahad, N., Hossen, M. J., & Ramanathan, T. T. (2025). Enhancing Alzheimer's disease detection: An explainable machine learning approach with ensemble techniques. Intelligence-BasedMedicine, 11, 100240. https://doi.org/10.1016/j.ibmed.2025.100240
[5] F.H. Saif, M.N. Al-Andoli, W. MohdExplainable AI for Alzheimer detection: a review of current methods and applicationsAppl Sci, 14 (22) (Nov. 2024), Article 10121, 10.3390/app142210121
[6] Tanveer, M., Richhariya, B., Khan, R. U., Rashid, A. H., Khanna, P., Prasad, M., & Lin, C. T. (2020). Machine learning techniques for the diagnosis of Alzheimer's disease: A review. ACM Transactions on MultimediaComputing, Communications, and Applications, 16(1s), Article 30, 1–35. https://doi.org/10.1145/3344998
[7] Malik, I., Iqbal, A., Gu, Y. H., & Al-antari, M. A. (2024). Deep Learning for Alzheimer's Disease Prediction: A Comprehensive Review. Diagnostics, 14(12), 1281. https://doi.org/10.3390/diagnostics14121281
[8] Song S, Li T, Lin W, Liu R and Zhang Y (2025) Application of artificial intelligence in Alzheimer'sdisease:abibliometricanalysis. Front. Neurosci. 19:1511350. doi: 10.3389/fnins.2025.1511350.
[9] Govindarajan, R., Thirunadanasikamani, K., Napa, K. K., Sathya, S., Senthil Murugan, J., & Chandi Priya, K. G. (2025). Development of an explainable machine learning model for Alzheimer's disease prediction using clinical and behavioural features. MethodsX, 15, 103491. https://doi.org/10.1016/j.mex.2025.103491
[10] Shahbaz M, Ali S, Guergachi A, Niazi A, Umer A. Classification of Alzheimer's Disease Using Machine Learning Techniques. In: Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies. Prague, CzechRepublic; 2019. p. 296-303. doi: 10.5220/0007949902960303.
[11] Vlontzou, M. E., Athanasiou, M., Dalakleidi, K. V., Nikita, K. S., &Dinov, I. D. (2025). A comprehensive interpretable machine learning framework for mild cognitive impairment and Alzheimer's disease diagnosis. Scientific Reports, 15, 8410. https://doi.org/10.1038/s41598-025-92577-6

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[12] Samad, A., &Aydı, E. S. (2024). Rapid Alzheimer's disease diagnosis using advanced artificial intelligence algorithms. International Journal of Innovative Science and Research Technology, 9(6), 1760–1766. https://doi.org/10.38124/ijisrt/IJISRT24JUN1915
[13] El Kharoua, R. (2024). Alzheimer's Disease Dataset [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/8668279
[14] Chawla, N. V., Bowyer, K. W., Hall, L. O., &Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
[15] Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. Journal of Statistical Software, 36(11), 1–13.
[16] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 4765–4774 (2017).
[17] Liu, H., Roeder, K., & Wasserman, L. (2010). Stability Approach to Regularization Selection (StARS) for high dimensional graphical models. Advances in Neural Information ProcessingSystems (NeurIPS), 23, 1432–1440.
[18] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7, 1–30.