scientific reports

Check for updates

OPEN

# An optimized machine learning framework integrating feature selection and Bayesian–Tree-Structured Parzen estimator hyperparameter optimization for alzheimer's disease prediction

Rajkumar Govindarajan[1], K. Thirunadanasikamani[1], Komal Kumar Napa[2], Haiter Lenin Allasi[3]✉, Mary Vasanthi Soosaimariyan[4] & C. Sheeja Herobin Rani[4]

**Alzheimer's disease (AD) is an increasing global health challenge due to its rising prevalence and the limited effectiveness of existing diagnostic approaches. Neuropathological changes often begin more than a decade before clinical symptoms emerge, yet most diagnostic methods detect the disease only after significant cognitive decline. Machine learning offers promising solutions for early prediction; however, predictive performance is often limited by suboptimal feature selection and hyperparameter optimization. This study proposes an integrated predictive framework that combines three feature selection methods—ReliefF, Boruta, and Recursive Feature Elimination with Cross-Validation (RFE(RFECV))—with probabilistic hyperparameter optimization using the Tree-Structured Parzen Estimator (TPE) and Bayesian Optimization using Gaussian Processes (BayesGP) for Random Forest classification. A clinical dataset comprising 2,149 subjects and 34 features was preprocessed using missing value imputation, ordinal encoding, Min–Max normalization, and SMOTE applied exclusively to the training data. Model performance was evaluated on an independent test set using bootstrap confidence intervals, McNemar's test, and calibration analysis. The Boruta + TPE + Random Forest configuration achieved the best overall performance with 95.04% accuracy, 93.13% F1-score, and an AUC of 98.41%, while the RFE(RFECV) + TPE model achieved the highest AUC (98.48%). These findings demonstrate that integrating robust feature selection with advanced hyperparameter optimization substantially improves predictive accuracy, model reliability, and interpretability, providing an effective framework for early Alzheimer's disease risk stratification and clinical decision support.**

**Keywords** Alzheimer's disease classification, Boruta feature selection, Recursive feature elimination, Tree-Structured Parzen Estimator, Bayesian optimization

Alzheimer's disease (AD) is one of the most prevalent neurodegenerative disorders and a major global public health challenge, imposing substantial emotional, social, and economic burdens on patients, caregivers, and healthcare systems[1]. It is characterized by progressive deterioration in cognitive function, memory, and functional independence, ultimately leading to severe disability and complete dependency[2]. Neuropathological evidence indicates that amyloid-β plaques and tau neurofibrillary tangles begin accumulating many years before the onset of clinical symptoms, highlighting the importance of early identification and risk stratification for improving patient outcomes[3,4]. With the aging global population, the number of individuals living with AD is

[1]Department of Computer Science and Engineering, St. Peter's Institute of Higher Education and Research, Avadi, Chennai, India. [2]Department of Artificial Intelligence and Data Science, Saveetha Engineering College, Chennai, Tamil Nadu, India. [3]School of Engineering and Applied Sciences, Kampala International University, Box 20000, Kampala, Uganda. [4]Department of Electronics and Communication Engineering, St Xavier's Catholic College of Engineering, Nagercoil, Tamilnadu, India. ✉email: drahlenin@kiot.edu.et; haiterlenina@gmail.com

projected to increase from approximately 50 million to more than 150 million by 2050, emphasizing the urgent need for scalable, accessible, and cost-effective diagnostic and predictive approaches[5].

Alzheimer's disease progresses from a preclinical stage through mild cognitive impairment (MCI) to severe dementia, with multiple phenotypic variants including early-onset and atypical forms affecting language or visuospatial function[6,7]. Current diagnosis relies primarily on cognitive assessments, neuroimaging, and biomarker analyses; however, these approaches are often expensive, time-consuming, and limited to specialized clinical settings, reducing their accessibility for routine early detection[8–10].

Machine learning (ML) has emerged as a promising approach for early AD prediction because of its ability to model complex nonlinear relationships within clinical and behavioral data[11–13]. Among various ML algorithms, Random Forest (RF) has demonstrated strong predictive performance, robustness, and suitability for heterogeneous medical datasets[14]. Nevertheless, predictive performance is often limited by high-dimensional feature spaces, class imbalance, and sensitivity to hyperparameter settings, which may lead to overfitting and reduced generalization[15,16]. Addressing these challenges requires effective feature selection to eliminate redundant variables and improve interpretability[17], together with robust hyperparameter optimization to maximize predictive performance on unseen data[18]. Conventional approaches such as filter-based ranking and grid search are often computationally inefficient and may not adequately explore the complex search space of clinical datasets[19].

To address these limitations, this study proposes an integrated predictive framework that combines three complementary feature selection techniques—ReliefF, Boruta, and Recursive Feature Elimination with Cross-Validation (RFE/RFECV)—to identify the most informative predictors of Alzheimer's disease[20–22]. Furthermore, two adaptive hyperparameter optimization strategies, namely the Tree-Structured Parzen Estimator (TPE) and Bayesian Optimization with Gaussian Processes (BayesGP), are employed to optimize the Random Forest classifier for improved predictive accuracy and stability[23,24]. The proposed framework aims to identify clinically relevant predictors, enhance classification performance through adaptive optimization, and comprehensively evaluate model robustness using cross-validation, bootstrap confidence interval estimation, and McNemar's statistical significance test. By integrating feature refinement and model optimization within a unified analytical pipeline, the proposed approach seeks to improve early-stage Alzheimer's disease prediction while supporting computational efficiency and practical clinical decision-making.

## Related work

Machine learning has been extensively investigated for Alzheimer's disease prediction, with studies employing diverse datasets comprising neuroimaging, clinical, cognitive, and behavioral information[25]. This diversity in input modalities has resulted in considerable variation in methodological designs and reported predictive performance. Classical machine learning algorithms, including Support Vector Machines, have achieved accuracy levels between 80 and 90% in neuroimaging-based classification tasks[26], whereas ensemble learning methods, particularly Random Forest classifiers, have demonstrated improved robustness and generalization on multimodal clinical datasets, frequently achieving accuracies exceeding 90%[27].

Despite these advances, most existing studies focus on optimizing either feature selection or hyperparameter tuning independently, rather than jointly addressing both aspects within a unified predictive framework[28]. Recent developments in artificial intelligence have facilitated the integration of demographic characteristics, medical history, lifestyle factors, cognitive assessments, and functional measures into comprehensive prediction models[29]. These computational approaches can identify complex nonlinear relationships that are often difficult to detect through conventional clinical assessment, thereby supporting earlier risk stratification and diagnosis[30,31]. However, Alzheimer's disease datasets frequently contain high-dimensional and heterogeneous features that introduce redundancy and increase computational complexity, while inappropriate hyperparameter settings may adversely affect model stability and generalization[32,33].

Feature selection therefore plays a crucial role in identifying clinically meaningful predictors while reducing model complexity[34]. In addition to conventional statistical approaches, nature-inspired optimization techniques such as the Whale Optimization Algorithm and Artificial Bee Colony Optimization have demonstrated promising performance in selecting informative feature subsets from complex search spaces[35,36]. Likewise, hyperparameter optimization has increasingly adopted metaheuristic algorithms, including Ant Colony Optimization and Bald Eagle Search, to efficiently explore high-dimensional parameter spaces and improve predictive performance with lower computational cost than exhaustive search methods[37–39].

Random Forest has emerged as a particularly effective classifier for Alzheimer's disease prediction because of its ability to model nonlinear relationships, reduce overfitting, estimate feature importance, and handle missing data efficiently[40]. Nevertheless, its predictive performance remains highly dependent on both the quality of the selected feature subset and appropriate hyperparameter configuration, highlighting the need for integrated optimization strategies[41].

Recent studies further demonstrate the potential of machine learning for Alzheimer's disease prediction. Kumar et al.[42] achieved approximately 87% accuracy using clinical and lifestyle variables, demonstrating the feasibility of cost-effective screening approaches. Mubonanyikuzo et al.[43] reported approximately 89% accuracy by integrating neuroimaging with clinical biomarkers, while Biswas and Gini[44] achieved 92% accuracy by combining cognitive assessments with genetic risk factors. Although these studies report encouraging predictive performance, they generally optimize either feature selection or hyperparameter tuning in isolation and do not systematically investigate their combined influence on model performance.

Consequently, few studies have comprehensively compared multiple feature selection techniques together with multiple hyperparameter optimization strategies within a unified Random Forest framework. As a result, the optimal combination of these complementary approaches for improving Alzheimer's disease prediction remains insufficiently explored[45]. The present study addresses this research gap by systematically evaluating

three complementary feature selection methods (ReliefF, Boruta, and RFECV) alongside two adaptive Bayesian optimization strategies (TPE and BayesGP) within a single integrated predictive framework.

## Method

This study presents an integrated machine learning framework for Alzheimer's disease (AD) prediction by combining advanced feature selection techniques with model-driven hyperparameter optimization within a unified analytical pipeline. The proposed workflow comprised all key stages of predictive model development, including data preprocessing, feature selection, model training, hyperparameter optimization, performance evaluation, and statistical validation. Three complementary feature selection techniques were investigated: ReliefF, a filter-based algorithm that estimates feature relevance based on nearest-neighbor instances; Boruta, a Random Forest-based wrapper method that identifies all relevant features; and Recursive Feature Elimination with Cross-Validation (RFE/RFECV), which iteratively removes less informative variables using cross-validation. To determine the optimal Random Forest configuration, two adaptive hyperparameter optimization approaches were employed: the Tree-Structured Parzen Estimator (TPE) and Bayesian Optimization with Gaussian Processes (BayesGP). Model performance was comprehensively evaluated using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC). Model robustness was further assessed through bootstrap-based 95% confidence intervals and statistical significance testing using McNemar's test.

All experiments were implemented in Python 3.11 using open-source scientific computing libraries. Data preprocessing and exploratory analysis were performed using NumPy[56] and Pandas[57], while preprocessing, Random Forest implementation, Recursive Feature Elimination with Cross-Validation (RFECV), model training, and performance evaluation were carried out using Scikit-learn[55]. Class imbalance in the training data was addressed using the Synthetic Minority Oversampling Technique (SMOTE)[53] implemented through the imbalanced-learn[58] library. Boruta feature selection was implemented using BorutaPy, whereas ReliefF was implemented using a Python implementation of the original ReliefF algorithm. Hyperparameter optimization was performed using Optuna[54] with the Tree-Structured Parzen Estimator sampler and scikit-optimize for Bayesian optimization based on Gaussian Processes. Statistical analyses, including McNemar's test, were conducted using statsmodels[59], and all visualizations were generated using Matplotlib[60].

### Dataset

This study utilized a publicly available AD dataset obtained from the Kaggle repository. The dataset represents the clinical and cognitive assessment records aggregated from multiple healthcare sources. A total of 2,149 patient records were included, each described by 34 attributes covering a broad range of clinical and personal factors. These features encompassed:

- Demographic details: age, gender, marital status, and educational attainment
- Lifestyle and behavioral characteristics: smoking status, alcohol consumption, physical activity, and sleep patterns
- Medical history: family history of Alzheimer's disease, cardiovascular disease, diabetes, depression, head trauma, and hypertension
- Cognitive and functional assessments: memory recall, orientation, problem-solving ability, and daily living activity performance

The diagnostic label classified individuals into either Alzheimer's disease (AD) or non-AD groups. The dataset exhibited class imbalance, with 1,389 non-AD cases and 760 AD cases, reflecting realistic disparities encountered in clinical populations. Recognizing this imbalance, SMOTE was applied during training to ensure fair model learning without altering the original test distribution.

Table 1 provides a complete listing of all 34 features, including their descriptions, clinical relevance, value ranges, and encoding schemes for categorical variables. This structured feature documentation ensures transparency and supports reproducibility for future studies. The comprehensive baseline characteristics is provided in Supplementary Table S1 online.

### Data preprocessing

The dataset comprises 2,149 patient records characterized by a diverse set of 34 attributes, which include demographic characteristics, lifestyle and behavioral factors, medical history related to neurodegeneration, cognitive evaluation indicators, and functional impairment measures. Initially, the dataset exhibited class imbalance, consisting of 1,389 non-AD cases and 760 AD cases. All patient identifiers were pre-anonymized, ensuring compliance with ethical data privacy standards.

A detailed description of all clinical, behavioral, and functional variables used in this study is provided in Table 1, including their medical relevance, encoding formats, and supported value ranges. Key attributes include age (60–90 years), gender, ethnicity, educational background, body mass index (BMI), smoking and alcohol habits, diet and sleep quality assessment scores, physical activity profiles, family history of AD, and comorbid medical conditions such as hypertension, diabetes, cardiovascular disorders, head injury history, and depression.

Missing values in numerical features were imputed using the median value strategy, while categorical variables were handled using most frequent category imputation. Subsequently, to ensure uniform scaling and eliminate biases introduced by differing feature ranges, Min–Max normalization was applied to all features using Eq. (1):

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

| Feature name | Description | Category |
|---|---|---|
| PatientID | Unique patient identifier | Administrative |
| Age | Age of the individual (years) | Demographic |
| Gender | Biological sex of the individual | Demographic |
| Ethnicity | Ethnic background | Demographic |
| EducationLevel | Highest level of education attained | Demographic |
| BMI | Body Mass Index (kg/m$^2$) | Lifestyle |
| Smoking | Smoking status (current/former/never) | Lifestyle |
| AlcoholConsumption | Frequency of alcohol intake | Lifestyle |
| PhysicalActivity | Level of regular physical exercise | Lifestyle |
| DietQuality | Quality of dietary habits | Lifestyle |
| SleepQuality | Self-reported sleep quality | Lifestyle |
| FamilyHistoryAlzheimers | Presence of Alzheimer's in first-degree relatives | Medical History |
| CardiovascularDisease | History of cardiovascular conditions | Medical History |
| Diabetes | Presence of diabetes | Medical History |
| Depression | History of depressive episodes | Medical History |
| HeadInjury | History of head trauma | Medical History |
| CholesterolTotal | Total cholesterol level (mg/dL) | Medical History |
| CholesterolLDL | LDL cholesterol level (mg/dL) | Medical History |
| CholesterolHDL | HDL cholesterol level (mg/dL) | Medical History |
| CholesterolTriglycerides | Triglyceride level (mg/dL) | Medical History |
| MMSE | Mini-Mental State Examination score | Cognitive Assessment |
| FunctionalAssessment | Functional ability assessment score | Functional Status |
| MemoryComplaints | Reported memory difficulties | Cognitive Assessment |
| BehavioralProblems | Presence of behavioral disturbances | Clinical Symptoms |
| ADL | Activities of Daily Living score | Functional Status |
| Confusion | Episodes of confusion | Clinical Symptoms |
| Disorientation | Disorientation to time/place | Clinical Symptoms |
| PersonalityChanges | Changes in personality traits | Clinical Symptoms |
| DifficultyCompletingTasks | Problems executing familiar tasks | Functional Status |
| Forgetfulness | Frequency of forgetting recent events | Cognitive Assessment |
| DoctorInCharge | Attending physician label (non-clinical) | Administrative |
| Diagnosis | Class label for Alzheimer's disease | Outcome Variable |

**Table 1**. Feature Descriptions and Categories.

where $x_{norm}$ is the normalized feature value, $x$ is the original value, $x_{min}$ and $x_{max}$ represent the minimum and maximum values of the feature, respectively.

To preserve diagnostic class proportionality, a stratified 70:30 split was used to divide the dataset into training and testing subsets using random_state = 44. The independent test set comprised 645 subjects, including 417 Non-AD and 228 AD cases, thereby maintaining the original class distribution. To mitigate the effect of class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied only to the training data to prevent information leakage into the test set, using k_neighbors = 6 and random_state = 44. For reproducibility, the same random_state = 44 was consistently used throughout the experimental pipeline, including the train–test split, Random Forest classifier initialization, Tree-Structured Parzen Estimator (TPE) optimization (Optuna sampler), and Bayesian Optimization with Gaussian Processes (BayesGP). This ensured deterministic data partitioning, model training, and hyperparameter optimization across all experiments. After SMOTE oversampling, the training dataset achieved balanced class distribution, consisting of approximately equal representations of AD and non-AD cases. This resulted in 1,507 training instances (753 non-AD and 754 AD samples), thus enhancing the model's capability to learn minority-class decision boundaries effectively.

## Feature selection

Feature selection plays a critical role in developing robust and generalizable machine learning models, particularly when handling high-dimensional clinical datasets. By identifying the most relevant predictors and eliminating redundant or noisy features, feature selection enhances classification performance, reduces computational overhead, and minimizes the risk of model overfitting[46]. In this study, three complementary feature selection approaches were implemented to systematically identify the most informative predictors associated with Alzheimer's disease: ReliefF, Boruta, and Recursive Feature Elimination with Cross-Validation (RFE/RFECV).

### ReliefF feature selection

ReliefF[50] is a distance-based filter method that estimates the relevance of each feature by measuring how well it distinguishes between instances of different classes. For each sample $x_i$, ReliefF identifies its nearest neighbor from the same class (hit) and from a different class (miss), based on Euclidean distance in feature space. The relevance weight $W(f)$ for a feature $f$ is updated iteratively as:

$$W(f) = W(f) - \frac{1}{m}\sum_{i=1}^{m} \mid x_{i,f} - Hit_{i,f} \mid + \frac{1}{m}\sum_{i=1}^{m} \mid x_{i,f} - Miss_{i,f} \mid \quad (2)$$

where $m$ is the number of sampled instances, $x_{i,f}$ is the value of feature $f$ for instance , $Hit_{i,f}$ is the nearest same-class neighbor, $Miss_{i,f}$ is the nearest opposite-class neighbor.

A positive $W(f)$ implies that the feature effectively differentiates between diagnostic classes, whereas a negative value suggests low or detrimental discriminatory contribution. Only features with $W(f) > 0$ were retained for modeling.

### Boruta feature selection

Boruta is a wrapper-based selection algorithm that leverages the Random Forest importance measure to evaluate feature relevance[51]. To assess significance, the method generates shadow features by randomly permuting each original feature, preserving its distribution while removing informative structure.

For each feature $f$, the importance score $I(f)$ is computed using Mean Decrease in Impurity (MDI):

$$I(f) = \sum_{t \in T_f} \Delta i(t) \quad (3)$$

where $T_f$ denotes all tree nodes where feature $f$ is used for splitting, $\Delta i\,(t)$ represents the impurity reduction at node $t$.

A feature is retained if:

$$I(f) > max\{I(f_{shadow})\} \quad (4)$$

This ensures that only features demonstrating statistically significant predictive contribution beyond randomness are selected. Unlike greedy elimination methods, Boruta preserves all relevant predictors, supporting interpretability in clinical decision-making.

### Recursive feature elimination with cross-validation (RFE/RFECV)

RFE is a wrapper method that recursively removes the least important feature based on model-specific importance rankings[52]. In this study, a Random Forest classifier was used as the evaluation estimator due to its robustness and ability to capture non-linear dependencies.

At each iteration, the feature importance ranking $r(f)$ is computed as:

$$r(f) = \frac{1}{\mid T_f \mid}\sum_{t \in T_f} \Delta i(t) \quad (5)$$

Features with the lowest ranking values are removed stepwise. To determine the optimal number of retained features, RFECV extends RFE with 10-fold stratified cross-validation, selecting the feature subset that maximizes the mean validation accuracy across the ten folds:

$$F^* = argmax Accuracy_{CV}(F_k) \quad (6)$$

where $F_k$ is the subset of features of size $k$.

All feature selection methods were applied only on the SMOTE-balanced training dataset to avoid information leakage. The resulting feature subsets were subsequently used for hyperparameter optimization and model development.

### Hyperparameter optimization

Tuning hyperparameters is a vital part of building effective machine learning models, since the choice of these settings has a direct impact on how well the model performs, how consistently it behaves across different data, and how well it generalizes to unseen cases[47]. In this study, two adaptive optimization strategies were employed to fine-tune the hyperparameters of the Random Forest classifier: Tree-Structured Parzen Estimator (TPE) and Bayesian Optimization[24] using Gaussian Processes (BayesGP). Both methods are probabilistic search strategies designed to efficiently explore the hyperparameter space while minimizing computational overhead.

The TPE algorithm models the hyperparameter search as a conditional probability distribution and iteratively prioritizes regions of the search space associated with high performance. This approach balances exploration and exploitation by probabilistically selecting and evaluating candidate hyperparameter configurations based on expected improvement.

Conversely, Bayesian Optimization (BayesGP) employs Gaussian Process regression to construct a surrogate model of the objective function. By defining an acquisition function that estimates the likelihood of improvement,

BayesGP selects hyperparameter sets that maximize predictive gain while reducing unnecessary evaluations. This makes it particularly suitable for computationally expensive search spaces.

For both optimization strategies, the Random Forest classifier's predictive accuracy under stratified 10-fold cross-validation served as the optimization objective. The following hyperparameters were tuned:

- Number of Trees (n_estimators): Controls the number of decision trees in the ensemble
- Maximum Depth (max_depth): Regulates tree complexity to balance flexibility and overfitting
- Minimum Samples Split (min_samples_split): Determines the minimum number of samples required to split a node
- Minimum Samples Leaf (min_samples_leaf): Specifies the minimum number of samples allowed at a terminal node
- Maximum Features (max_features): Limits the number of features considered when selecting the best split

The hyperparameter search was executed only on the feature-selected, SMOTE-balanced training dataset to ensure unbiased evaluation. The optimal hyperparameter configurations obtained from TPE and BayesGP were subsequently used to train final Random Forest models for external testing.

## Tree-Structured Parzen Estimator (TPE)

The Tree-Structured Parzen Estimator (TPE)[23] is a sequential model-based hyperparameter optimization method that replaces the traditional search over the objective function $f(\theta)$ with a search over probabilistic models of hyperparameters. Instead of modeling $p(y \mid \theta)$, TPE directly models the inverse distribution $p(\theta \mid y)$, which allows more efficient sampling from promising hyperparameter regions.

TPE partitions the observations of the objective function into good and poor configurations based on a given threshold $y^*$:

$$p(\theta \mid y) = \begin{cases} l(\theta), if y < y^* \\ g(\theta), if y \geq y^* \end{cases} \tag{7}$$

where $l(\theta)$ is the density of hyperparameter values yielding better-than-threshold performance, $g(\theta)$ is the density for poorer outcomes.

The algorithm selects new hyperparameters $\theta^*$ that maximize the ratio:

$$\theta^* = argmax g \frac{l(\theta)}{g(\theta)} \tag{8}$$

This ratio biases exploration toward hyperparameter regions historically associated with better model performance.

The expected improvement (EI) objective optimized by TPE is:

$$EI(\theta) \propto \left( \gamma + \frac{1-\gamma}{\frac{g(\theta)}{l(\theta)}} \right)^{-1} \tag{9}$$

where $\gamma$ controls the quantile cutoff ($\gamma = 0.15$) in this study.

In this study, TPE was employed to optimize the Random Forest hyperparameters by minimizing the validation error over 10-fold cross-validation. Table 2 depicts the algorithm for optimizing Random Forest by TPE-Based hyperparameter.

## Bayesian optimization using Gaussian processes (BayesGP)

Bayesian Optimization[24] treats hyperparameter tuning as a probabilistic optimization problem, where the goal is to find:

$$\theta^* = argmax f(\theta) \tag{10}$$

where $f(\theta)$ is the unknown model performance metric, and $\theta$ denotes the hyperparameter vector.

A Gaussian Process (GP) is used as a surrogate model to approximate $f(\theta)$. A GP defines a prior over functions:

$$f(\theta) \sim GP(m(\theta), k(\theta, \theta')) \tag{11}$$

where $m(\theta)$ is the mean function, $k(\theta, \theta')$ is the kernel (covariance) function such as Radial Basis Function (RBF).

Given observed performance values $\mathcal{D} = \{\theta_i, f(\theta_i)\}_{i=1}^{n}$, the GP posterior predictive distribution for a new point $\theta^*$ is:

$$\mu(\theta^*) = k^T K^{-1} y \tag{12}$$

$$\sigma^2(\theta^*) = k(\theta^*, \theta^*) - k^T K^{-1} k \tag{13}$$

```
Input:
   Training dataset Xtrain, ytrain
   Feature selection method (Boruta)
   Number of trials = 40

Output:
   Optimized Random Forest model M*
   Optimal hyperparameters θ*

1: Define the hyperparameter search space:
      n_estimators ∈ [100, 600]
      max_depth ∈ [4, 24]
      min_samples_split ∈ [2, 20]
      min_samples_leaf ∈ [1, 10]
      max_features ∈ {sqrt, log2, None}
2: Apply Boruta feature selection:
      XtrainFS ← Boruta(Xtrain)
      ▷ Retains the most informative features
3: Initialize:
      best_score ← 0
      best_params ← NULL
4: for trial = 1 to 40 do
5:      Sample hyperparameter set θ using TPE
6:      Train Random Forest with θ
7:      Evaluate model using
       10-fold stratified cross-validation
8:      score ← Mean CV accuracy
9:      Update TPE posterior using (θ, score)
10:     if score > best_score then
11:           best_score ← score
12:           best_params ← θ
13:     end if
14: end for
15: Train the final Random Forest model using best_params
16: Evaluate the optimized model on the independent test set
17: Compute:
      Accuracy
      Precision
      Recall
      F1-score
      AUC
18: Return M* and best_params
```

**Table 2**. Algorithm for Optimizing of Random Forest by TPE-Based Hyperparameter.

where $K$ is the covariance matrix of past observations, $k$ is the covariance vector between $\theta^*$ and training points.

To select the next hyperparameter candidate, Bayesian Optimization maximizes an acquisition function such as Expected Improvement (EI):

$$EI(\theta^*) = (\mu(\theta^*) - f_{best})\Phi(Z) + \sigma(\theta^*)\phi(Z) \tag{14}$$

where $f_{best}$ is the best observed performance, $Z = \frac{\mu(\theta^*) - f_{best}}{\sigma(\theta^*)}$, $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative and probability density functions of the standard normal distribution.

The BayesGP optimization was performed exclusively on the feature-selected, SMOTE-balanced training data using 10-fold stratified cross-validation during hyperparameter optimization. The independent test set was completely excluded from the optimization process, thereby preventing information leakage and ensuring unbiased hyperparameter selection. Table 3 presents the BayesGP-based hyperparameter optimization procedure for the Random Forest classifier.

Input:
Training dataset Xtrain, ytrain
Feature selection method (Boruta)
Number of optimization iterations = 35

Output:
Optimized Random Forest model M*
Optimal hyperparameters θ*

1: Define the hyperparameter search space:
n_estimators ∈ [100, 600]
max_depth ∈ [4, 24]
min_samples_split ∈ [2, 20]
min_samples_leaf ∈ [1, 10]
max_features ∈ {sqrt, log2, None}
2: Apply Boruta feature selection:
XtrainFS ← Boruta(Xtrain)
▷ Retains the most informative features
3: Initialize:
Gaussian Process (GP) surrogate model
Acquisition function
n_calls ← 35
4: for iteration = 1 to 35 do
5: Select the next hyperparameter set θ by maximizing
the acquisition function using the current GP surrogate model
6: Train Random Forest with θ
7: Evaluate the model using
10-fold stratified cross-validation
8: score ← Mean CV accuracy
9: Update the Gaussian Process surrogate model
with the new observation (θ, score)
10: end for
11: Select the optimal hyperparameter set θ*
corresponding to the highest cross-validation accuracy
12: Train the final Random Forest model using θ*
13: Evaluate the optimized model on the independent test set
14: Compute:
Accuracy
Precision
Recall
F1-score
AUC
15: Return M* and θ*

**Table 3**. Algorithm for Optimizing Random Forest Using BayesGP-Based Hyperparameter Optimization.

## AD prediction using random forest

Random Forest (RF) was selected as the primary predictive model in this study due to its robustness, interpretability, and proven effectiveness in handling heterogeneous and high-dimensional clinical data[48]. RF is an ensemble learning algorithm that constructs multiple decision trees during the training process and combines their outputs to produce the final classification result. In the case of classification tasks such as Alzheimer's disease prediction, the final class label is determined using majority voting across all trees in the ensemble.

The predictive output of the Random Forest classifier can be expressed as:

$$y^{\wedge} = mode\{h_1(x), h_2(x), \ldots, h_T(x)\} \tag{15}$$

where $y^{\wedge}$ is the predicted class label, $h_t(x)$ represents the prediction of the $t^{th}$ decision tree, $T$ is the total number of trees in the forest.

By aggregating the predictions of multiple trees, RF reduces the risk of overfitting, improves generalization performance, and is particularly effective when working with noisy, incomplete, or non-linearly separable medical datasets.

To ensure optimal performance, the Random Forest hyperparameters were tuned using the Tree-Structured Parzen Estimator (TPE) and Bayesian Optimization with Gaussian Processes (BayesGP). The optimized hyperparameters included the number of decision trees, tree depth, and splitting criteria, which directly influenced model complexity and decision boundary behavior.

Once optimized, the final RF model was trained using the feature-selected and SMOTE-balanced training dataset, and performance was evaluated on the independent test set. Evaluation metrics included accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC) to comprehensively assess prediction quality, class discrimination capability, and clinical reliability.

The selection of Random Forest in this study was based on its strong historical performance in Alzheimer's disease research, ability to manage complex variable interactions, and its capacity to produce interpretable feature importance rankings for clinical insights[49]. The integration of advanced hyperparameter optimization strategies further enhanced its predictive capability, making it well-suited for early-stage Alzheimer's disease risk prediction.

### Performance metrics

To ensure a reliable and unbiased assessment of model performance, a stratified hold-out validation strategy was implemented, where the dataset was divided into 70% for training and 30% for testing, maintaining the original class distribution in both splits. Additionally, during hyperparameter optimization, a 10-fold stratified cross-validation scheme was applied on the training set only, ensuring stability of parameter tuning and preventing information leakage into the test set.

Model performance was evaluated using multiple standard performance metrics, including Accuracy, Precision, Recall, and F1-score, which quantify different aspects of classification effectiveness. These metrics were computed using the following equations:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{16}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{17}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{18}$$

$$F1 - Score = 2 \times \frac{Precision * Recall}{Precision + Recall} \tag{19}$$

where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

To further evaluate the discriminative ability of the model, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was computed, providing insight into the classifier's separability between Alzheimer's and non-Alzheimer's cases.

To quantify statistical reliability, 95% confidence intervals for Accuracy, Precision, Recall, F1-score, and AUC were estimated using bootstrap resampling (1,000 iterations) applied to the predictions of the independent held-out test set after completion of model training, thereby preserving complete isolation of the test data during feature selection, SMOTE balancing, and hyperparameter optimization. Additionally, McNemar's test was performed to assess whether performance differences between the best-performing models were statistically significant, supporting meaningful comparative evaluation. The complete Alzheimer's disease prediction workflow, including preprocessing, feature selection, hyperparameter optimization, and final evaluation, is illustrated in Fig. 1.

## Results

This section presents the experimental outcomes obtained through the integration of feature selection techniques and hyperparameter optimization strategies for Alzheimer's disease prediction. The primary objective of the study was to evaluate how different feature selection and parameter tuning combinations influence the performance of the Random Forest classifier in identifying Alzheimer's disease cases.

### Identification of significant features for Alzheimer's disease prediction

Three feature selection techniques were employed to identify the most informative predictors associated with Alzheimer's disease. These selected features provide clinically meaningful insights that can support early diagnosis, risk assessment, and targeted intervention strategies in healthcare settings.

The ReliefF method identified a subset of 15 features that demonstrated strong discriminative power in distinguishing Alzheimer's disease cases from controls. The selected features included Age, Education Level, Diet Quality, Physical Activity, Sleep Quality, Family History of Alzheimer's, Cardiovascular Disease, Head Injury, Hypertension, Depression, MMSE Score, Functional Assessment Score, Disorientation. Memory Complaints, and Difficulty Completing Tasks. Using this subset, the Random Forest model achieved 92.37% ± 2.14% accuracy under 10-fold cross-validation.

The Boruta method, which uses Random Forest-based importance ranking, identified 13 key features as strongly relevant to Alzheimer's disease. These included Age, Education Level, Cardiovascular Disease,

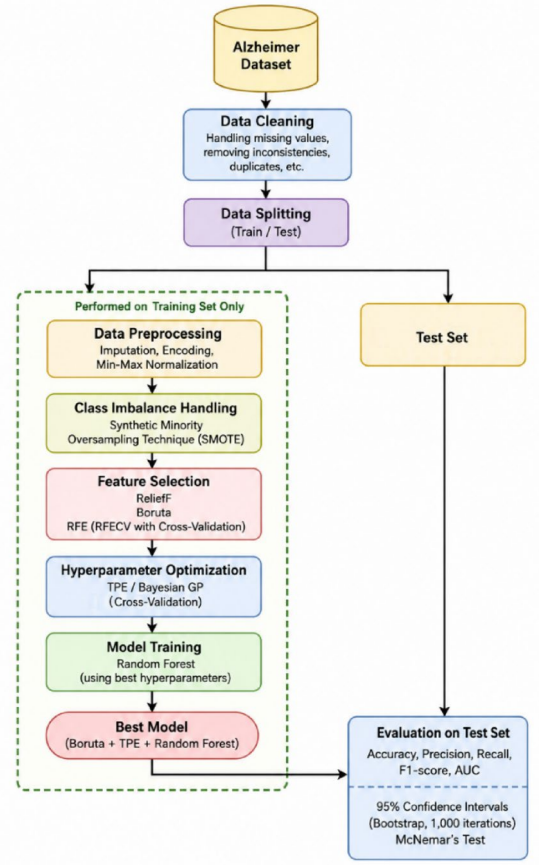



**Fig. 1**. Alzheimer disease prediction workflow

Diet Quality, Family History of Alzheimer's, Head Injury, Hypertension, Diabetes, Depression, MMSE Score, Functional Assessment, Memory Complaints, and Forgetfulness. This subset demonstrated the best performance in the study when paired with TPE-optimized Random Forest, achieving 95.04% ± 1.71% accuracy and 98.41% ± 0.87% AUC on the held-out test set.

The Recursive Feature Elimination with Cross-Validation (RFE/RFECV) approach selected 18 features based on model-driven iterative elimination. This subset included all Boruta-selected features, along with additional behavioral and cognitive indicators such as ADL Score, Confusion, Personality Changes, and Behavioural Problems. Figure 2 illustrates the overlap among the features selected by ReliefF, Boruta, and RFE(RFECV). When combined with Bayesian optimization, this feature subset achieved 95.04% ± 1.63% accuracy and 98.28% ± 0.85% AUC, matching the best performance in classification accuracy while providing enhanced class separation capability.

## Hyperparameter optimization results

Hyperparameter optimization plays a predominant role in enhancing model performance, particularly for ensemble-based approaches such as Random Forest, where parameters such as tree depth, ensemble size, and node splitting thresholds can substantially influence classification outcomes. In this study, two probabilistic optimization strategies were employed: Tree-Structured Parzen Estimator (TPE) and Bayesian Optimization with Gaussian Processes (BayesGP). These methods were selected to overcome the limitations associated with conventional grid and random search techniques, including excessive computational overhead, risk of suboptimal convergence, and limited adaptability in high-dimensional search spaces.

Table 4 presents the optimized hyperparameters derived using each feature selection–optimizer combination, along with the total optimization time (minutes) required for one complete hyperparameter optimization run, including all cross-validation folds. Among ReliefF-based feature subsets, TPE produced a more compact model with 132 trees and moderate depth (max_depth = 13) in 1.02 min, whereas BayesGP favored a larger ensemble (581 trees, depth = 19), requiring 1.17 min.

For the Boruta-selected features, TPE again produced a relatively efficient model with 362 trees and shallow leaf constraints, completing optimization in 1.57 min. In contrast, BayesGP identified a slightly smaller but deeper model (414 trees, max_depth = 15), taking 1.66 min. Similarly, for RFE(RFECV)-based feature subsets, TPE identified a configuration with 300 trees and greater depth (21), while BayesGP selected an equivalent number of trees but with shallower depth (15). Notably, BayesGP required slightly more optimization time (1.75 min) compared to TPE (1.49 min).

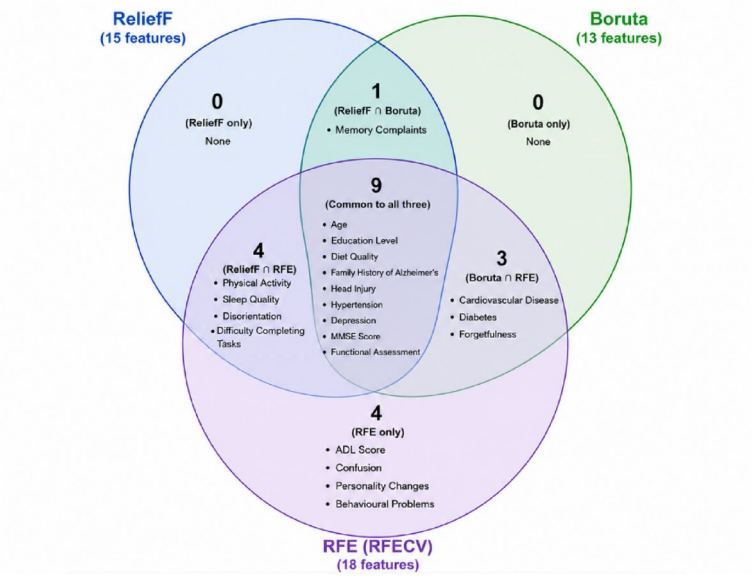



**Fig. 2**. Venn diagram illustrating the overlap among the features selected by ReliefF, Boruta, and RFE(RFECV)

| Configuration (FS + Optimizer) | n_estimators | max_depth | min_samples_split | min_samples_leaf | Total time (mins) |
|---|---|---|---|---|---|
| ReliefF + TPE | 132 | 13 | 14 | 6 | 1.02 |
| ReliefF + BayesGP | 581 | 19 | 2 | 3 | 1.17 |
| Boruta + TPE | 362 | 20 | 5 | 2 | 1.57 |
| Boruta + BayesGP | 414 | 15 | 3 | 2 | 1.66 |
| RFE(RFECV) + TPE | 300 | 21 | 4 | 2 | 1.49 |
| RFE(RFECV) + BayesGP | 300 | 15 | 3 | 2 | 1.75 |

**Table 4**. Optimized Random Forest hyperparameters obtained using TPE and BayesGP algorithms for the Alzheimer's dataset.

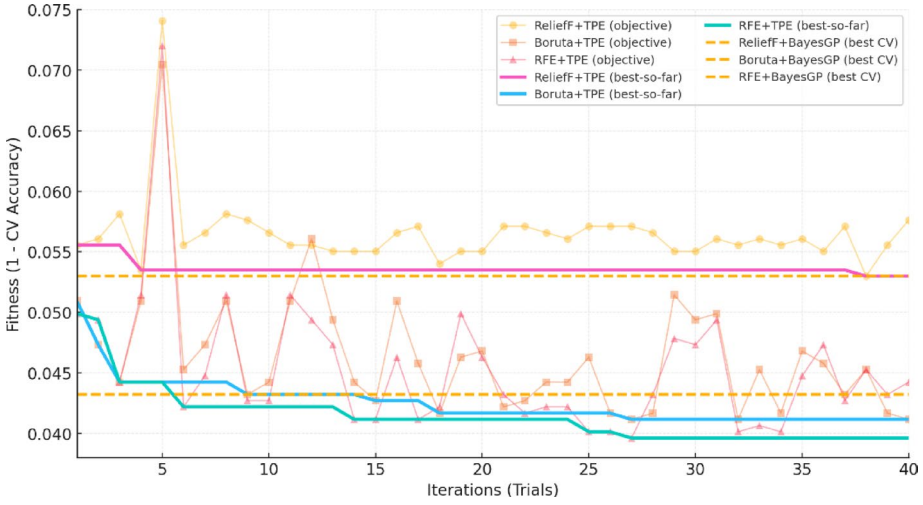



**Fig. 3**. Fitness Function Convergence Over Iterations (TPE trials with BayesGP best-CV references)

The convergence behavior of the hyperparameter optimization process is illustrated in Fig. 3, which depicts the progression of the fitness function across optimization iterations for both the Tree-Structured Parzen Estimator (TPE) and Bayesian Optimization using Gaussian Processes (BayesGP). In this study, the fitness function was defined as $1 - \overline{CV\,Accuracy}$, where $\overline{CV\,Accuracy}$ represents the mean accuracy across all cross-

validation folds. Consequently, minimizing the fitness function corresponds to maximizing the average cross-validation performance.

The convergence patterns indicate that BayesGP generally achieved stable fitness values within fewer optimization iterations, demonstrating rapid convergence through Gaussian Process surrogate modeling. In contrast, TPE continued exploring the hyperparameter search space over additional iterations, progressively refining the candidate solutions before reaching comparable or slightly improved final fitness values for some feature selection configurations. BayesGP exhibited faster convergence, whereas TPE provided a more exploratory search behavior during the optimization process.

Across all feature selection strategies, the best-performing configuration was Boruta + TPE, which converged to the lowest final fitness score (corresponding to 95.04% accuracy and 98.41% AUC). While BayesGP matched this accuracy for RFE(RFECV), its convergence required slightly more iterations and computational time.

Time comparison showed that TPE completed hyperparameter optimization approximately 10–18% faster than BayesGP, with average optimization times of 1.02–1.57 min for TPE and 1.17–1.75 min for BayesGP, respectively. This demonstrates TPE's stronger exploration–exploitation balance and reduced computational overhead.

## Statistical validation

To assess the robustness and statistical reliability of the proposed Alzheimer's disease prediction framework, McNemar's test and 10-fold cross-validation were performed. McNemar's test was used to compare the classification performance of the two best-performing configurations, namely Boruta + TPE + Random Forest and RFE(RFECV) + BayesGP + Random Forest.

The test produced a McNemar statistic of $\chi^2 = 0.050$ with $p = 0.823$, indicating no statistically significant difference between the two optimized Random Forest configurations. This finding suggests that although minor differences exist in precision, recall, and AUC, both models exhibit statistically comparable classification performance.

Furthermore, 10-fold cross-validation demonstrated consistent and stable performance of the proposed framework. The best-performing Boruta + TPE + Random Forest model achieved an average accuracy of 95.04% ± 1.71%, precision of 91.18% ± 3.56%, recall of 95.18% ± 2.73%, F1-score of 93.13% ± 2.42%, and AUC of 98.41% ± 0.87%, confirming the robustness and generalizability of the proposed approach. Detailed McNemar contingency matrices and pairwise statistical comparisons between all evaluated classifiers are provided in the Supplementary Information (Tables S4 and S5).

## Experimental results

The optimal hyperparameters identified through TPE-based optimization for the Random Forest classifier were n_estimators = 362, max_depth = 20, min_samples_split = 5, and min_samples_leaf = 2. These parameters were selected based on the highest cross-validation score obtained using the Boruta-selected feature subset (13 features), providing an effective balance between model complexity and generalization performance.

Table 5 summarizes the performance of the optimized Random Forest classifier across all feature selection methods (ReliefF, Boruta, and RFE(RFECV)) using both TPE and Bayesian Gaussian Process (BayesGP) optimization. Overall, the Boruta + TPE + Random Forest configuration achieved the best overall performance, recording the highest accuracy (95.04% ± 1.71%), recall (95.18% ± 2.73%), F1-score (93.13% ± 2.42%), and AUC (98.41% ± 0.87%). The RFE(RFECV) + BayesGP model achieved the same highest accuracy (95.04% ± 1.63%), whereas RFE(RFECV) + TPE produced the highest AUC (98.48% ± 0.85%), indicating excellent discriminative capability. Although the ReliefF-based models demonstrated competitive classification performance, they consistently produced lower accuracy and AUC values than the Boruta- and RFE(RFECV)-based configurations. These findings indicate that Boruta and RFE(RFECV) generated more informative feature subsets for Alzheimer's disease prediction, while both TPE and BayesGP effectively optimized the Random Forest classifier. Figure 4 further illustrates the comparative performance of all feature selection and optimization strategies. Comparisons of model performance with and without SMOTE, along with class-level performance metrics, are provided in Supplementary Tables S2 and S3 online, respectively.

The confusion matrices obtained for all feature selection and optimization configurations is depicted in Fig. 5. Across the models, Boruta combined with TPE and BayesGP demonstrated consistently strong discriminative capability, yielding high True Negative (Non-AD correctly classified) and True Positive (AD correctly classified) counts. Specifically, the Boruta + TPE configuration produced 396 true negatives and 217 true positives, while

| Feature selection | Optimizer | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) | AUC (%) |
|---|---|---|---|---|---|---|
| ReliefF | TPE | 94.57 ± 1.71 | 93.27 ± 3.80 | 91.23 ± 3.72 | 92.24 ± 2.80 | 94.61 ± 2.28 |
| | BayesGP | 94.88 ± 1.69 | 93.72 ± 3.51 | 91.67 ± 3.90 | 92.68 ± 2.84 | 94.34 ± 2.27 |
| Boruta | TPE | 95.04 ± 1.71 | 91.18 ± 3.56 | 95.18 ± 2.73 | 93.13 ± 2.42 | 98.41 ± 0.87 |
| | BayesGP | 94.57 ± 1.57 | 90.04 ± 3.42 | 95.18 ± 2.48 | 92.54 ± 2.30 | 98.35 ± 0.81 |
| RFE(RFECV) | TPE | 94.73 ± 1.63 | 90.76 ± 3.21 | 94.74 ± 2.76 | 92.70 ± 2.58 | 98.48 ± 0.85 |
| | BayesGP | 95.04 ± 1.63 | 92.98 ± 3.18 | 92.98 ± 3.46 | 92.98 ± 2.68 | 98.28 ± 0.85 |

**Table 5**. Performance comparison of Random Forest models using different feature selection and hyperparameter optimization strategies for Alzheimer's disease prediction.

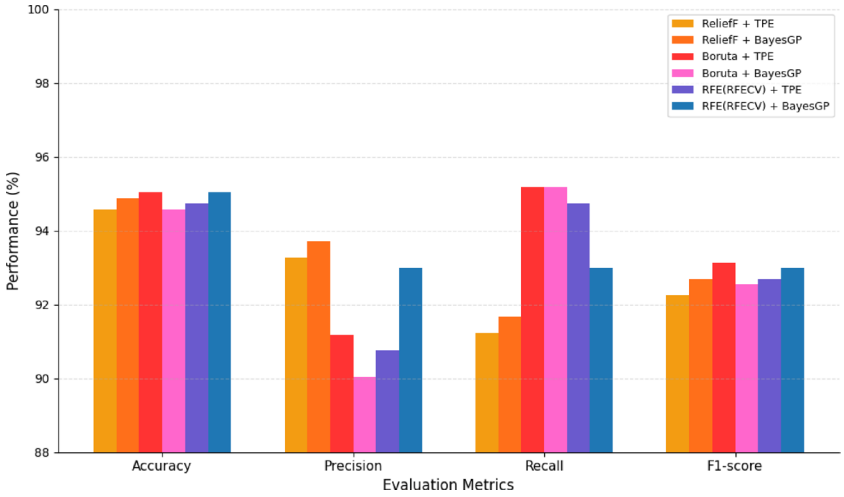



**Fig. 4**. Performance comparison

Boruta + BayesGP achieved 393 true negatives and 217 true positives, indicating reliable sensitivity and specificity in Alzheimer's disease detection.

The RFE(RFECV)-based models also delivered strong results, with RFE + BayesGP correctly identifying 401 Non-AD and 212 AD cases, and RFE + TPE yielding 395 true negatives and 216 true positives, demonstrating competitive performance and robust class discrimination.

In comparison, the ReliefF-based models showed slightly higher misclassification rates, particularly in AD detection, where ReliefF + TPE and ReliefF + BayesGP resulted in lower true positive counts (208–209) and slightly higher false negatives. However, they still maintained strong true negative performance (> 402 correct Non-AD classifications), reflecting effective detection of non-diseased cases.

## Feature importance

The feature importance ranking derived from the Boruta + TPE optimized Random Forest model highlights a set of clinically meaningful predictors associated with AD diagnosis. The feature importance is shown in Fig. 6. Functional Assessment (0.235) emerged as the most influential feature, followed closely by MMSE (0.198) and ADL (0.178). These top-ranked cognitive and daily-function assessment measures collectively represent core indicators used in clinical diagnostic evaluations, reaffirming their relevance in differentiating Alzheimer's cases from non-affected individuals.

Symptoms related to cognitive behavior and memory, including Memory Complaints (0.152) and Behavioral Problems (0.145), also showed substantial predictive power, indicating their strong association with early cognitive decline and neurobehavioral changes commonly observed in Alzheimer's progression. The presence of PatientID (0.089) among selected features suggests dataset-level identity-linked variance, which was retained due to its statistical relevance in the model despite offering limited clinical interpretability. Interestingly, Diabetes appeared as a lower-ranked yet retained feature, reflecting the well-documented link between metabolic disorders and increased neurodegenerative risk. This aligns with recent clinical findings suggesting that systemic health comorbidities can contribute to or accelerate cognitive impairment.

To further assess the clinical utility of the most influential predictors, an additional experiment was conducted using only the five highest-ranked features identified by the Boruta + TPE + Random Forest model. The simplified model achieved comparable predictive performance, and the detailed results are provided in the Supplementary Information (See Supplementary Figs. S1, S2 online and Supplementary Table S6 online).

## Comparison with machine learning classifiers

Table 6 presents a comprehensive comparison of classification performance across multiple machine learning classifiers combined with three feature selection strategies (ReliefF, Boruta, and RFE(RFECV)) and their corresponding model configurations. The results demonstrate that Random Forest, SVM, and Stacking Ensemble consistently outperform Naïve Bayes, Logistic Regression, and KNN across all evaluation metrics. Identical performance values observed for some classifier–optimizer combinations indicate that TPE and BayesGP converged to the same optimal hyperparameter configuration, resulting in equivalent predictive performance.

Among the Random Forest models, the Boruta + TPE + RF configuration achieved the highest performance with an accuracy of 0.9504 [0.9333–0.9674], a balanced precision–recall profile, and the highest AUC (0.9841 [0.9745–0.9919]). This indicates that Boruta's ability to retain stable, biologically meaningful features enhances predictive discriminability when paired with efficient hyperparameter tuning via TPE. In comparison, SVM models also performed strongly. Particularly, RFE(RFECV) + BayesGP + SVM achieved 0.9194 [0.8992–0.9395] accuracy and 0.9725 [0.9616–0.9826] AUC, closely approaching Random Forest performance while maintaining superior precision in distinguishing Alzheimer's cases.

The Stacking Ensemble models showed consistently high performance across all feature selection strategies, with Boruta + TPE + Stacking achieving 0.9411 [0.9225–0.9597] accuracy and 0.9774 [0.9655–0.9885] AUC.

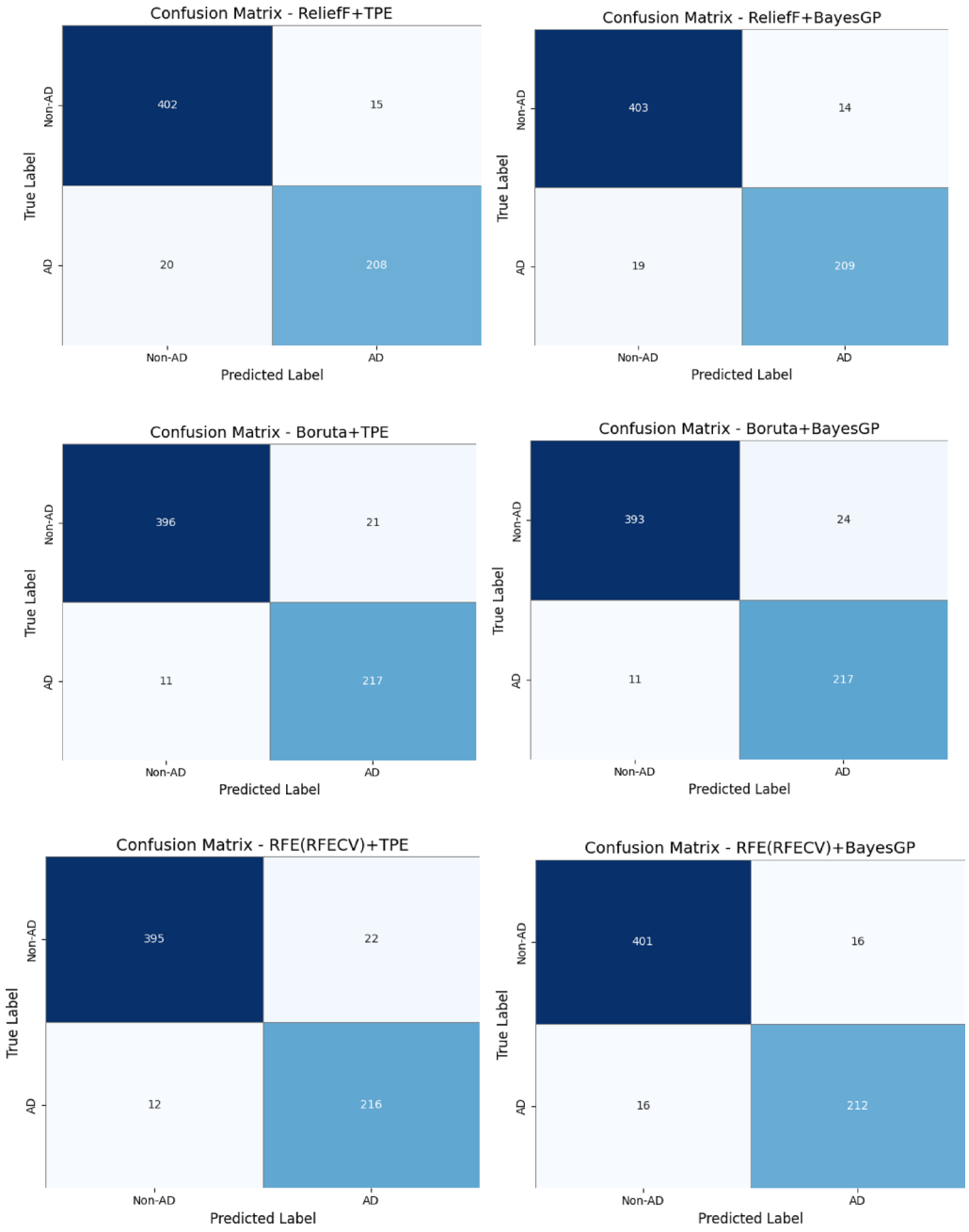



**Fig. 5**. Comparison of Confusion matrices

This demonstrates the benefit of integrating complementary classifiers for improved stability and robustness in prediction. Conversely, Naive Bayes and Logistic Regression models exhibited relatively lower performance, with accuracy values clustering between 0.80 and 0.83 and reduced precision–recall consistency. These results reflect the limited ability of linear-probabilistic classifiers to capture non-linear interactions among clinical, cognitive, and behavioral predictors in Alzheimer's disease.

Figure 7 illustrates the ROC curves for all evaluated feature selection and optimization configurations. The results clearly show that Boruta + TPE and RFE(RFECV) + TPE achieved the highest discriminative performance, with AUC values of 0.984 and 0.985, respectively, closely followed by their Bayesian GP counterparts (AUC ≈ 0.983–0.984). In comparison, the ReliefF-based configurations exhibited comparatively lower performance (AUC ≈ 0.943–0.946), indicating that feature selection strategy played a critical role in classification accuracy.

The ROC curves further demonstrate that the Boruta-based models, particularly Boruta + TPE, consistently achieved higher sensitivity across all specificity thresholds, especially in the high-specificity region, which is clinically important for minimizing false-positive Alzheimer's diagnoses. Similarly, the RFE(RFECV) + TPE and BayesGP configurations exhibited strong balance between true positive and false positive rates, confirming their reliability and robustness.

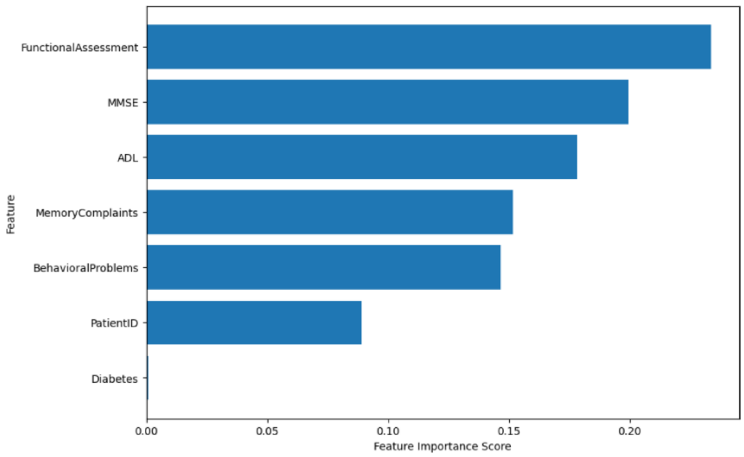



**Fig. 6**. Feature importance ranking of the Boruta + TPE + Random Forest model based on Mean Decrease in Impurity (MDI). The feature importance scores were computed from the optimized Random Forest classifier trained using the Boruta-selected feature subset. Higher importance scores indicate a greater contribution of individual features to Alzheimer's disease classification.

| Configuration | Accuracy (95% CI) | Precision (95% CI) | Recall (95% CI) | F1-score (95% CI) | AUC (95% CI) |
|---|---|---|---|---|---|
| ReliefF + TPE + RF | 0.9457 [0.9287, 0.9628] | 0.9327 [0.8945, 0.9625] | 0.9123 [0.8750, 0.9469] | 0.9224 [0.8946, 0.9466] | 0.9461 [0.9233, 0.9659] |
| ReliefF + BayesGP + RF | 0.9488 [0.9318, 0.9659] | 0.9372 [0.9010, 0.9662] | 0.9167 [0.8816, 0.9506] | 0.9268 [0.9015, 0.9499] | 0.9434 [0.9193, 0.9641] |
| **Boruta + TPE + RF** | **0.9504 [0.9333, 0.9674]** | **0.9118 [0.8739, 0.9451]** | **0.9518 [0.9227, 0.9772]** | **0.9313 [0.9053, 0.9536]** | **0.9841 [0.9745, 0.9919]** |
| Boruta + BayesGP + RF | 0.9457 [0.9271, 0.9628] | 0.9004 [0.8611, 0.9375] | 0.9518 [0.9227, 0.9775] | 0.9254 [0.9004, 0.9471] | 0.9835 [0.9735, 0.9916] |
| RFE(RFECV) + TPE + RF | 0.9473 [0.9302, 0.9628] | 0.9076 [0.8683, 0.9427] | 0.9474 [0.9178, 0.9737] | 0.9270 [0.9014, 0.9499] | 0.9848 [0.9752, 0.9922] |
| RFE(RFECV) + BayesGP + RF | 0.9504 [0.9333, 0.9659] | 0.9298 [0.8945, 0.9602] | 0.9298 [0.8969, 0.9595] | 0.9298 [0.9043, 0.9511] | 0.9828 [0.9736, 0.9905] |
| ReliefF + TPE + NaiveBayes | 0.8016 [0.7690, 0.8310] | 0.6866 [0.6289, 0.7431] | 0.8070 [0.7577, 0.8590] | 0.7419 [0.6978, 0.7836] | 0.8657 [0.8366, 0.8945] |
| Boruta + TPE + NaiveBayes | 0.8202 [0.7907, 0.8465] | 0.7014 [0.6434, 0.7527] | 0.8553 [0.8100, 0.8987] | 0.7708 [0.7261, 0.8093] | 0.8861 [0.8585, 0.9113] |
| RFE(RFECV) + TPE + NaiveBayes | 0.8279 [0.7984, 0.8527] | 0.7082 [0.6499, 0.7584] | 0.8728 [0.8294, 0.9132] | 0.7819 [0.7383, 0.8169] | 0.8915 [0.8650, 0.9152] |
| ReliefF + TPE + LogReg | 0.8326 [0.8016, 0.8621] | 0.7239 [0.6655, 0.7783] | 0.8509 [0.8009, 0.8967] | 0.7823 [0.7366, 0.8215] | 0.8961 [0.8698, 0.9214] |
| ReliefF + BayesGP + LogReg | 0.8310 [0.8000, 0.8605] | 0.7196 [0.6625, 0.7748] | 0.8553 [0.8054, 0.9000] | 0.7816 [0.7366, 0.8202] | 0.8955 [0.8692, 0.9210] |
| Boruta + TPE + LogReg | 0.8295 [0.8000, 0.8605] | 0.7169 [0.6593, 0.7710] | 0.8553 [0.8075, 0.9018] | 0.7800 [0.7366, 0.8189] | 0.9006 [0.8762, 0.9252] |
| Boruta + BayesGP + LogReg | 0.8295 [0.8000, 0.8605] | 0.7169 [0.6593, 0.7710] | 0.8553 [0.8075, 0.9018] | 0.7800 [0.7366, 0.8189] | 0.9007 [0.8762, 0.9254] |
| RFE(RFECV) + TPE + LogReg | 0.8326 [0.8031, 0.8636] | 0.7190 [0.6627, 0.7765] | 0.8640 [0.8170, 0.9067] | 0.7849 [0.7430, 0.8242] | 0.8998 [0.8757, 0.9250] |
| RFE(RFECV) + BayesGP + LogReg | 0.8233 [0.7922, 0.8527] | 0.7096 [0.6513, 0.7621] | 0.8465 [0.7964, 0.8933] | 0.7720 [0.7299, 0.8117] | 0.9025 [0.8783, 0.9271] |
| ReliefF + TPE + SVM | 0.8667 [0.8419, 0.8946] | 0.8008 [0.7480, 0.8541] | 0.8289 [0.7773, 0.8792] | 0.8147 [0.7763, 0.8535] | 0.9167 [0.8907, 0.9415] |
| ReliefF + BayesGP + SVM | 0.8698 [0.8450, 0.8977] | 0.8051 [0.7532, 0.8598] | 0.8333 [0.7854, 0.8837] | 0.8190 [0.7813, 0.8584] | 0.9179 [0.8926, 0.9422] |
| Boruta + TPE + SVM | 0.9039 [0.8822, 0.9256] | 0.8517 [0.8018, 0.8950] | 0.8816 [0.8384, 0.9224] | 0.8664 [0.8318, 0.8990] | 0.9615 [0.9474, 0.9745] |
| Boruta + BayesGP + SVM | 0.9039 [0.8822, 0.9256] | 0.8517 [0.8018, 0.8950] | 0.8816 [0.8384, 0.9224] | 0.8664 [0.8318, 0.8990] | 0.9616 [0.9475, 0.9745] |
| RFE(RFECV) + TPE + SVM | 0.9163 [0.8946, 0.9349] | 0.8537 [0.8057, 0.8979] | 0.9211 [0.8869, 0.9533] | 0.8861 [0.8554, 0.9129] | 0.9729 [0.9618, 0.9830] |
| RFE(RFECV) + BayesGP + SVM | 0.9194 [0.8992, 0.9395] | 0.8607 [0.8144, 0.9048] | 0.9211 [0.8865, 0.9537] | 0.8898 [0.8596, 0.9178] | 0.9725 [0.9616, 0.9826] |
| ReliefF + TPE + KNN | 0.8326 [0.8047, 0.8620] | 0.7256 [0.6747, 0.7835] | 0.8465 [0.7957, 0.8945] | 0.7814 [0.7383, 0.8215] | 0.9003 [0.8743, 0.9252] |
| ReliefF + BayesGP + KNN | 0.8326 [0.8047, 0.8620] | 0.7256 [0.6747, 0.7835] | 0.8465 [0.7957, 0.8945] | 0.7814 [0.7383, 0.8215] | 0.9003 [0.8743, 0.9252] |
| Boruta + TPE + KNN | 0.8992 [0.8744, 0.9209] | 0.8221 [0.7721, 0.8659] | 0.9123 [0.8733, 0.9456] | 0.8649 [0.8289, 0.8953] | 0.9503 [0.9322, 0.9658] |
| Boruta + BayesGP + KNN | 0.8992 [0.8744, 0.9209] | 0.8221 [0.7721, 0.8659] | 0.9123 [0.8733, 0.9456] | 0.8649 [0.8289, 0.8953] | 0.9503 [0.9322, 0.9658] |
| RFE(RFECV) + TPE + KNN | 0.8992 [0.8760, 0.9225] | 0.8147 [0.7623, 0.8622] | 0.9254 [0.8889, 0.9580] | 0.8665 [0.8326, 0.8974] | 0.9627 [0.9477, 0.9757] |
| RFE(RFECV) + BayesGP + KNN | 0.8992 [0.8760, 0.9225] | 0.8147 [0.7623, 0.8622] | 0.9254 [0.8889, 0.9580] | 0.8665 [0.8326, 0.8974] | 0.9627 [0.9477, 0.9757] |
| ReliefF + TPE + Stacking | 0.9426 [0.9240, 0.9597] | 0.9170 [0.8770, 0.9520] | 0.9211 [0.8869, 0.9528] | 0.9190 [0.8908, 0.9434] | 0.9497 [0.9281, 0.9702] |
| ReliefF + BayesGP + Stacking | 0.9426 [0.9240, 0.9597] | 0.9170 [0.8770, 0.9520] | 0.9211 [0.8869, 0.9528] | 0.9190 [0.8908, 0.9434] | 0.9497 [0.9282, 0.9703] |
| Boruta + TPE + Stacking | 0.9411 [0.9225, 0.9597] | 0.8893 [0.8468, 0.9267] | 0.9518 [0.9211, 0.9783] | 0.9195 [0.8920, 0.9436] | 0.9774 [0.9655, 0.9885] |
| Boruta + BayesGP + Stacking | 0.9411 [0.9225, 0.9597] | 0.8893 [0.8468, 0.9267] | 0.9518 [0.9211, 0.9783] | 0.9195 [0.8920, 0.9436] | 0.9773 [0.9651, 0.9884] |
| RFE(RFECV) + TPE + Stacking | 0.9426 [0.9240, 0.9597] | 0.8930 [0.8515, 0.9306] | 0.9518 [0.9213, 0.9774] | 0.9214 [0.8956, 0.9449] | 0.9755 [0.9621, 0.9877] |
| RFE(RFECV) + BayesGP + Stacking | 0.9426 [0.9240, 0.9597] | 0.8930 [0.8515, 0.9306] | 0.9518 [0.9213, 0.9774] | 0.9214 [0.8956, 0.9449] | 0.9779 [0.9658, 0.9887] |

**Table 6**. Comparison with ML classifiers. Significant values are in bold.

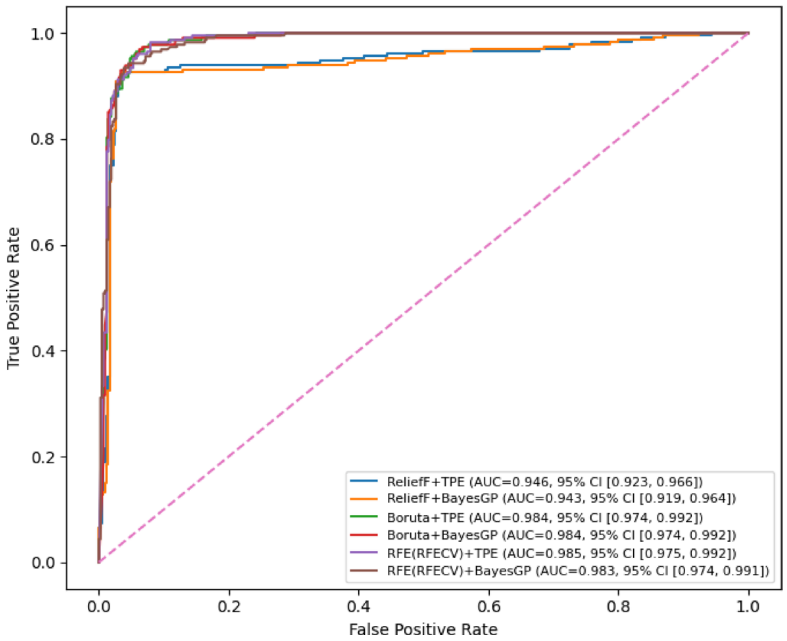



**Fig. 7**. Receiver Operating Characteristic (ROC) curves of all evaluated Random Forest configurations on the independent test set. The curves illustrate the classification performance of the ReliefF-, Boruta-, and RFE(RFECV)-based feature selection methods combined with TPE and BayesGP hyperparameter optimization. The reported AUC values and 95% confidence intervals were computed from predictions on the held-out test set. No macro- or micro-averaging and no ROC confidence/error bands are shown, as each curve represents the binary classification performance of an individual model.

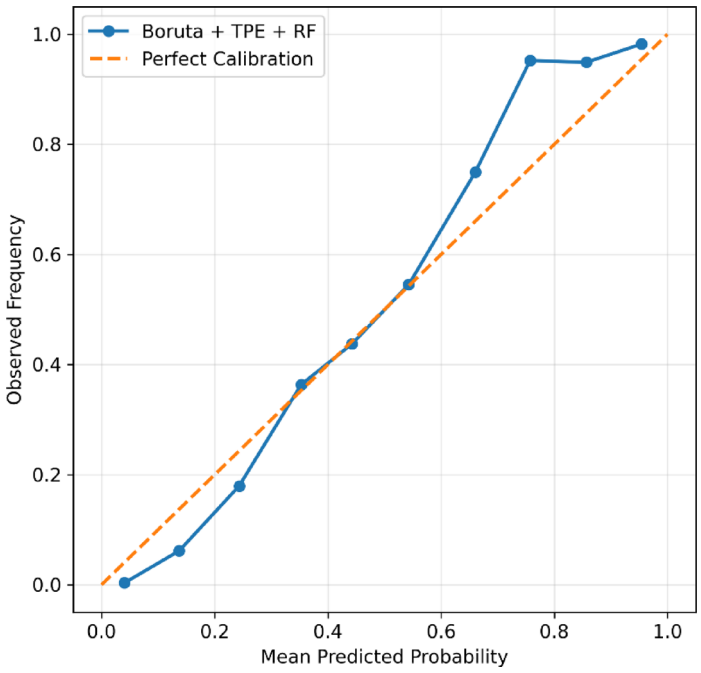



**Fig. 8**. Calibration curve for Boruta + TPE + RF

## Calibration analysis of the best model

To further evaluate the reliability of the predicted probabilities, the calibration performance of the best-performing Boruta + TPE + Random Forest model was assessed using a calibration curve and the Brier score. As illustrated in Fig. 8, the calibration curve closely follows the ideal diagonal reference line across most probability intervals, indicating good agreement between the predicted probabilities and the observed outcomes. Although

| Model | Brier score |
|---|---|
| Boruta + TPE + RF | 0.0472 |

**Table 7**. Brier score of the best-performing model.

| Optimization / Feature Selection Method | Total Task Time (mins) | Efficiency Gain |
|---|---|---|
| Grid Search (Baseline) | 90 min | Baseline |
| TPE Hyperparameter Optimization | 1.50 min (avg) | ≈ 98% reduction |
| Bayesian GP Optimization | 1.86 min (avg) | ≈ 98% reduction |
| ReliefF Feature Selection | 0.8 min | ≈ 89% reduction |
| Boruta Feature Selection | 2.4 min | ≈ 73% reduction |
| RFE (RFECV) Feature Selection | 3.2 min | Baseline (Slowest FS) |
| Traditional Full Pipeline | ~ 140 min | Baseline |
| Proposed Hybrid FS + TPE/BayesGP Pipeline | ~ 26 min | ≈ 81% reduction |

**Table 8**. Computational efficiency comparison.

| Study & Year | Dataset | Method | Accuracy (%) | AUC |
|---|---|---|---|---|
| Li et al.[61] | ADNI | XGBoost | 91.0 | – |
| Rani et al.[62] | OASIS | Random Forest | 94.03 | – |
| Hashmi et al.[63] | ADNI | Explainable XGBoost | 94.3 | 98.2 |
| Lei[64] | ADNI | Multimodal Deep Learning (AD-Diff) | 94.36 (1-year ADNI) | – |
| Emrah Aslan and Yildirim Özüpak[65] | ADNI | K-nearest neighbor regression | 97.33 | – |
| Proposed | OASIS | Boruta + TPE + Random Forest | 95.56 | 98.02 |
| Proposed | Kaggle Clinical AD Dataset | Boruta + TPE + Random Forest | 95.04 | 98.41 |

**Table 9**. Comparison of the proposed method with recent Alzheimer's disease prediction studies.

slight deviations are observed at higher predicted probability ranges, the overall trend demonstrates satisfactory probability calibration. Furthermore, the model achieved a Brier score of 0.0472 (Table 7), indicating a low probabilistic prediction error and well-calibrated confidence estimates. The combination of a high discriminative ability (AUC = 98.41%) and a low Brier score confirms that the proposed Boruta + TPE + Random Forest model not only distinguishes Alzheimer's disease cases effectively but also produces reliable probability estimates suitable for clinical decision-support applications.

## Analysis of computational efficiency

Table 8 summarizes the computational efficiency of the proposed framework in comparison to traditional exhaustive search-based approaches. Conventional grid-search-based hyperparameter tuning required approximately 90 min, whereas the proposed TPE and Bayesian GP optimization strategies reduced this time to 1.50 min and 1.86 min, respectively, representing an ≈98% reduction in computational cost.

Similarly, for feature selection, RFE required the highest execution time, while ReliefF achieved the fastest performance with an ≈89% time reduction, followed by Boruta with ≈73% reduction. When integrated into the complete model development pipeline, the proposed hybrid framework reduced total processing time from ~ 140 to ~ 26 min, demonstrating an ≈81% improvement in overall computational efficiency.

## Analysis of cross validation

To ensure that the observed performance was not due to random sampling variations and to assess the generalization stability of the proposed model, a stratified 10-fold cross-validation repeated 10 times (total of 100 evaluation cycles) was conducted. The Boruta + TPE + RF model, which emerged as the best-performing configuration in the study, demonstrated highly stable predictive performance across all repeated folds. The model achieved a mean accuracy of 95.04 ± 1.71, precision of 91.18 ± 3.56, recall of 95.18 ± 2.73, and F1-score of 93.13 ± 2.42, with an AUC of 98.41 ± 0.87.

## Comparison with recent Alzheimer's disease prediction studies

Table 9 compares the proposed framework with recent Alzheimer's disease prediction studies conducted on publicly available datasets. On the Kaggle Clinical AD Dataset, the proposed Boruta + TPE + Random Forest framework achieved 95.04% accuracy and the highest AUC (98.41%) among the compared methods reporting AUC values. To further assess its generalizability, the same framework was additionally evaluated on the OASIS Longitudinal dataset[62], where it achieved 95.56% accuracy and an AUC of 98.02% using the same

preprocessing and evaluation protocol. Detailed experimental results, including the classification report, 95% bootstrap confidence intervals, ROC curve, confusion matrix, and calibration analysis, are available in the online Supplementary Information (Supplementary Tables S7–S8 and Supplementary Figs. S3–S5).

Compared with the studies of Li et al.[61] (91.0%), Rani et al.[62] (94.03%), Lei[64] (94.36% on the 1-year ADNI dataset), and Hashmi et al.[63] (94.3% accuracy and 98.2% AUC), the proposed framework demonstrated superior or comparable predictive performance across both public clinical datasets. Although Emrah Aslan and Yildirim Özüpak[65] reported a higher classification accuracy (97.33%) on the ADNI dataset, the proposed framework consistently achieved high discriminative capability, with AUC values exceeding 98% on both the Kaggle Clinical AD and OASIS Longitudinal datasets. These findings demonstrate that the proposed feature selection and hyperparameter optimization framework is robust, generalizable, and competitive with recent state-of-the-art approaches for Alzheimer's disease prediction.

## Discussion

This study presented an integrated machine learning framework for Alzheimer's disease (AD) prediction by combining feature selection with probabilistic and surrogate-model-based hyperparameter optimization. The proposed framework systematically evaluated three feature selection techniques—ReliefF, Boruta, and Recursive Feature Elimination with Cross-Validation (RFE(RFECV))—together with two hyperparameter optimization strategies, the Tree-Structured Parzen Estimator (TPE) and Bayesian Optimization using Gaussian Processes (BayesGP), for Random Forest classification. Among all evaluated configurations, the Boruta + TPE + Random Forest model achieved the best overall predictive performance while maintaining computational efficiency.

The feature importance analysis demonstrated that Functional Assessment, MMSE, ADL, Memory Complaints, and Behavioral Problems were the strongest predictors of Alzheimer's disease, consistent with established clinical knowledge regarding cognitive decline and functional impairment. The inclusion of modifiable lifestyle factors such as Diet Quality and Physical Activity further highlights the potential value of preventive interventions and personalized risk assessment, supporting recent evidence that lifestyle and cognitive reserve contribute to disease progression.

The comparative analysis showed that feature selection substantially influenced predictive performance. Boruta produced the most clinically informative feature subset by preserving relevant cognitive, behavioral, and clinical variables while minimizing redundant features. ReliefF generated comparatively smaller feature subsets with slightly lower predictive accuracy, whereas RFE(RFECV) achieved competitive performance at the expense of higher computational complexity due to recursive feature elimination. These findings emphasize that selecting an appropriate feature selection strategy is critical for balancing predictive accuracy, interpretability, and computational efficiency in clinical prediction models.

Hyperparameter optimization further enhanced model performance. Although BayesGP exhibited faster convergence during optimization, TPE consistently identified more effective hyperparameter configurations, resulting in superior predictive performance across most feature selection strategies. The optimized Random Forest models indicated that deeper tree structures with balanced leaf constraints effectively captured the complex interactions among cognitive, lifestyle, and clinical variables, demonstrating the importance of systematic hyperparameter optimization for improving model generalization.

The statistical validation further supports the reliability of the proposed framework. Bootstrap confidence intervals demonstrated low variability across repeated resampling, while McNemar's test indicated no statistically significant difference between the two best-performing Random Forest configurations. Furthermore, calibration analysis of the Boruta + TPE + Random Forest model yielded a low Brier score (0.0472), confirming that the predicted probabilities were well calibrated and suitable for clinical decision support.

The proposed integration of Boruta feature selection, TPE-based hyperparameter optimization, and Random Forest classification provides a robust, interpretable, and clinically applicable framework for Alzheimer's disease prediction. The findings demonstrate that careful feature selection and hyperparameter optimization substantially improve predictive performance while supporting future translation of machine learning-based decision support systems into clinical practice.

## Limitations

Although the proposed framework demonstrated excellent predictive performance, several limitations should be acknowledged. First, the study was conducted using a single publicly available clinical dataset, which may not fully capture the demographic, geographic, and clinical heterogeneity of Alzheimer's disease. Consequently, external validation using independent, multi-centre cohorts is required before routine clinical deployment.

Second, the dataset is cross-sectional and contains a binary diagnostic outcome, limiting the evaluation of disease progression and longitudinal risk prediction. Future studies should validate the proposed framework using larger longitudinal datasets to improve temporal prediction and model generalizability.

Third, although the framework was rigorously validated using stratified cross-validation, bootstrap confidence intervals, calibration analysis, and statistical significance testing, additional external validation remains necessary to minimize potential dataset-specific bias.

Finally, the proposed model relies exclusively on structured clinical, behavioral, and lifestyle variables and does not incorporate neuroimaging, genetic information (e.g., APOE ε4 genotype), or cerebrospinal fluid biomarkers (amyloid-β and tau proteins), which are widely recognized as important biomarkers for early Alzheimer's disease diagnosis. Future research will focus on integrating multimodal data sources, and validating the proposed framework in prospective clinical settings to further enhance predictive performance, interpretability, and clinical applicability.

## Conclusion and future work

This study presented an integrated machine learning framework for Alzheimer's disease prediction by combining three complementary feature selection methods (ReliefF, Boruta, and RFE(RFECV)) with two advanced hyperparameter optimization strategies (TPE and BayesGP) for Random Forest classification. Among all evaluated configurations, the Boruta + TPE + Random Forest model achieved the best overall performance, attaining 95.04% accuracy, 93.13% F1-score, and an AUC of 98.41%, while demonstrating robust performance through cross-validation, confidence interval estimation, statistical significance testing, and calibration analysis.

The proposed framework identified Functional Assessment, MMSE, ADL, Memory Complaints, and Behavioral Problems as the most influential predictors of Alzheimer's disease, highlighting the combined importance of cognitive, functional, behavioral, and lifestyle-related factors in early disease prediction. Furthermore, the comparative analyses demonstrated that systematic feature selection and hyperparameter optimization substantially improved predictive performance compared with conventional machine learning configurations.

Overall, the proposed methodology provides a robust, interpretable, and computationally efficient framework for Alzheimer's disease risk prediction using routinely collected clinical data, making it suitable for clinical decision support and early screening applications. Future work will focus on validating the framework using independent multi-center and longitudinal cohorts, integrating multimodal information such as neuroimaging and biomarker data, and investigating parsimonious prediction models based on the highest-ranked clinical features to further enhance predictive performance, interpretability, and clinical applicability.

## Data availability

The datasets used and analyzed during the current study are available from https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset. All code and materials used in this study are available from the corres ponding author upon reasonable request.

Received: 27 February 2026; Accepted: 13 August 2026
Published online: 17 August 2026

## References

1. Memudu, A. E., Olukade, B. A. & Alex, G. S. Neurodegenerative diseases: Alzheimer's disease. *Integr. Neuroimag., Computat. Neurosci., Artif. Intell.*, https://doi.org/10.1201/9781032711102-8 (2024).
2. Botella Lucena, P. & Heneka, M. T. Inflammatory aspects of Alzheimer's disease. *Acta Neuropathol.* **148**(1), 31. https://doi.org/10.1007/s00401-024-02790-2 (2024).
3. Thal, D. R., Murray, M. E. & Dickson, D. W. Alzheimer's disease (Aβ and tau proteinopathy). In *Greenfield's Neuropathology 10e Set* 1012–1041 (CRC Press, 2024). https://doi.org/10.1201/9781003389699-32.
4. Almeida, Z. L., Vaz, D. C. & Brito, R. M. Morphological and molecular profiling of amyloid-β species in Alzheimer's pathogenesis. *Mol. Neurobiol.* **62**(4), 4391–4419. https://doi.org/10.1007/s12035-024-04543-4 (2025).
5. Pandey, S. K. et al. Inductive and transfer learning-based hybrid model techniques for accurate and automated diagnosis of neurological diseases. *Brain Behav.* **15**(8), e70788. https://doi.org/10.1002/brb3.70788 (2025).
6. Singh, N. A. et al. Atypical Alzheimer's disease: new insights into an overlapping spectrum between the language and visual variants. *J. Neurol.* **271**(6), 3571–3585. https://doi.org/10.1007/s00415-024-12297-1 (2024).
7. Putcha, D. et al. Heterogeneous clinical phenotypes of sporadic early-onset Alzheimer's disease: a neuropsychological data-driven approach. *Alzheimer's Res. Therapy* **17**(1), 38. https://doi.org/10.1186/s13195-025-01689-8 (2025).
8. Hassan, Y. M., Wanas, A., Ali, A. A. & El-Sayed, W. M. Integrating artificial intelligence with nanodiagnostics for early detection and precision management of neurodegenerative diseases. *J. Nanobiotechnol.* **23**(1), 668. https://doi.org/10.1186/s12951-025-03719-x (2025).
9. Wankhede, P. R. et al. Artificial intelligence for noninvasive health diagnostics. *ACS Sens.* **10**(11), 8217–8255. https://doi.org/10.1021/acssensors.5c03171 (2025).
10. Choi, H. K., Chen, M., Goldston, L. L. & Lee, K. B. Extracellular vesicles as nanotheranostic platforms for targeted neurological disorder interventions. *Nano Converg.* **11**(1), 19. https://doi.org/10.1186/s40580-024-00426-5 (2024).
11. Kaur, I. & Sachdeva, R. Prediction models for early detection of Alzheimer: Recent trends and future prospects. *Arch. Comput. Meth. Eng.* **32**(6), 3565–3592. https://doi.org/10.1007/s11831-025-10246-3 (2025).
12. Nazir, A., Assad, A., Hussain, A. & Singh, M. Alzheimer's disease diagnosis using deep learning techniques: Datasets, challenges, research gaps and future directions. *Int. J. Syst. Assur. Eng. Manag.* https://doi.org/10.1007/s13198-024-02441-5 (2024).
13. Soladoye, A. A., Aderinto, N., Osho, D. & Olawade, D. B. Explainable machine learning models for early Alzheimer's disease detection using multimodal clinical data. *Int. J. Med. Inform.* https://doi.org/10.1016/j.ijmedinf.2025.106093 (2025).
14. Shaikh, T. A., Islam, T. U., Mir, S. R., Namgail, T. & Gulzar, I. U. H. Deep ensemble learning for intelligent healthcare computing: A case study of Alzheimer's disease. *Int. J. Next-Gener. Comput.* **15**, 107. https://doi.org/10.47164/ijngc.v15i2.1475 (2024).
15. Dhanka, S., Sharma, A., Kumar, A., Maini, S. & Vundavilli, H. Advancements in hybrid machine learning models for biomedical disease classification using integration of hyperparameter-tuning and feature selection methodologies: A comprehensive review. *Arch. Comput. Meth. Eng.* **33**(1), 289–324. https://doi.org/10.1007/s11831-025-10309-5 (2026).
16. Wilson, A. & Anwar, M. R. The future of adaptive machine learning algorithms in high-dimensional data processing. *Int. Trans. Artif. Intell.* **3**(1), 97–107. https://doi.org/10.33050/italic.v3i1.656 (2024).
17. Borah, K. et al. A review on advancements in feature selection and feature extraction for high-dimensional NGS data analysis. *Funct. Integr. Genomics* **24**(5), 139. https://doi.org/10.1007/s10142-024-01415-x (2024).
18. Zheng, H. et al. Learning from models beyond fine-tuning. *Nat. Mach. Intell.* **7**(1), 6–17. https://doi.org/10.1038/s42256-024-00961-0 (2025).
19. García Andreu, A., Costa, M., & Pastor, O. (2025, June). Feature Selection in Medical Imaging: A Comprehensive Review. In International Conference on Advanced Information Systems Engineering (pp. 142–154). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-94931-9_12
20. Zarei, S., Saffar, M., Shalbaf, R., Hassani Abharian, P., & Shalbaf, A. (2025). Explainable hierarchical machine-learning approaches for multimodal prediction of conversion from mild cognitive impairment to Alzheimer's disease. Physical and Engineering Sciences in Medicine, 1–19. https://doi.org/10.1007/s13246-025-01618-x
21. Haridas, N. T., Sanchez-Bornot, J. M., McClean, P. L., Wong-Lin, K., Alzheimer's Disease Neuroimaging Initiative (ADNI). Autoencoder imputation of missing heterogeneous data for Alzheimer's disease classification. *Healthc. Technol. Lett.* **11**(6), 452–460. https://doi.org/10.1049/htl2.12091 (2024).

22. Ahmed, R. et al. A novel integrated logistic regression model enhanced with recursive feature elimination and explainable artificial intelligence for dementia prediction. *Healthc. Anal.* **6**, 100362. https://doi.org/10.1016/j.health.2024.100362 (2024).
23. Wang, J., Ding, J. E., Liu, F., Kallioniemi, E., Wang, S., Tsai, W. X., & Yang, A. C. (2025). Flexible and explainable graph analysis for EEG-based alzheimer's disease classification. arXiv:2504.01329. https://doi.org/10.48550/arXiv.2504.01329
24. Yang, K., Liu, L. & Wen, Y. The impact of Bayesian optimization on feature selection. *Sci. Rep.* **14**(1), 3948. https://doi.org/10.1038/s41598-024-54515-w (2024).
25. Halkiopoulos, C., Gkintoni, E., Aroutzidis, A. & Antonopoulou, H. Advances in neuroimaging and deep learning for emotion detection: A systematic review of cognitive neuroscience and algorithmic innovations. *Diagnostics* **15**(4), 456. https://doi.org/10.3390/diagnostics15040456 (2025).
26. Jia, Y. L. et al. Application of machine learning techniques in the diagnostic approach of PTSD using MRI neuroimaging data: A systematic review. *Heliyon* https://doi.org/10.1016/j.heliyon.2024.e28559 (2024).
27. Haq, I. U. et al. Advancements in medical radiology through multimodal machine learning: A comprehensive overview. *Bioengineering* **12**(5), 477. https://doi.org/10.3390/bioengineering12050477 (2025).
28. da Silva Bandeira, J., & de Araújo Fagundes, R. A. (2025). Enhancing Alzheimer's Disease Diagnosis: Insights from MLP and 1D CNN Models. In Simpósio Brasileiro de Sistemas de Informação (SBSI) (pp. 359–368). SBC. https://doi.org/10.5753/sbsi.2025.246492
29. Saeed, D. K. & Nashwan, A. J. Harnessing artificial intelligence in lifestyle medicine: Opportunities, challenges, and future directions. *Cureus* https://doi.org/10.7759/cureus.85580 (2025).
30. Athar, M. Potentials of artificial intelligence in familial hypercholesterolemia: Advances in screening, diagnosis, and risk stratification for early intervention and treatment. *Int. J. Cardiol.* **412**, 132315. https://doi.org/10.1016/j.ijcard.2024.132315 (2024).
31. Srivastava, V., Kumar, R., Wani, M. Y., Robinson, K. & Ahmad, A. Role of artificial intelligence in early diagnosis and treatment of infectious diseases. *Infect. Dis.* **57**(1), 1–26. https://doi.org/10.1080/23744235.2024.2425712 (2025).
32. Abdalrada, A. S. Comprehensive review of machine learning approaches for Alzheimer's disease diagnosis and prognosis. *Int. J. Ethical AI Appl.* **1**(2), 47–56. https://doi.org/10.64229/q996hq21 (2025).
33. Malakouti, S. M., Menhaj, M. B. & Suratgar, A. A. Applying Grid search, random search, bayesian optimization, genetic algorithm, and particle swarm optimization to fine-tune the hyperparameters of the ensemble of ML models enhances its predictive accuracy for mud loss. *Res Squ* https://doi.org/10.21203/rs.3.rs-5187887/v1 (2024).
34. Saim, M. M., & Ammor, H. (2024, April). Comparative study of feature selection algorithms for cardiovascular disease prediction with artificial neural networks. In International Conference on Smart Applications and Data Analysis (pp. 218–229). Cham: Springer Nature, Switzerland. https://doi.org/10.1007/978-3-031-77040-1_16
35. Selvarajan, S. A comprehensive study on modern optimization techniques for engineering applications. *Artif. Intell. Rev.* **57**(8), 194. https://doi.org/10.1007/s10462-024-10829-9 (2024).
36. Akram, F., Aziz, S., Khan, N. A., Faizi, S. A., & Raza, K. (2024). Integrating artificial bee colony algorithms for deep learning model optimization: A comprehensive review. Solving with Bees: Transformative Applications of Artificial Bee Colony Algorithm, 73–102. https://doi.org/10.1007/978-981-97-7344-2_5
37. Chellappan, D. & Rajaguru, H. Machine learning meets meta-heuristics: Bald eagle search optimization and red deer optimization for feature selection in type II diabetes diagnosis. *Bioengineering* **11**(8), 766. https://doi.org/10.3390/bioengineering11080766 (2024).
38. Soladoye, A. A. et al. Enhancing Alzheimer's disease prediction using random forest: A novel framework combining backward feature elimination and ant colony optimization. *Curr. Res. Transl. Med.* **73**(4), 103526. https://doi.org/10.1016/j.retram.2025.103526 (2025).
39. Saha, R., Mukherjee, A., Bal, A., Chakrabarti, A. & Mandal, S. N. A novel non-invasive handwriting-based machine learning approach for early detection of Alzheimer's disease. *Arab. J. Sci. Eng.* **51**(5), 6433–6464. https://doi.org/10.1007/s13369-025-10702-z (2026).
40. Rimal, Y., Sharma, N. & Alsadoon, A. The accuracy of machine learning models relies on hyperparameter tuning: Student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms. *Multimed. Tools Appl.* **83**(30), 74349–74364. https://doi.org/10.1007/s11042-024-18426-2 (2024).
41. Bakır, H. & Ceviz, Ö. Empirical enhancement of intrusion detection systems: A comprehensive approach with genetic algorithm-based hyperparameter tuning and hybrid feature selection. *Arab. J. Sci. Eng.* **49**(9), 13025–13043. https://doi.org/10.1007/s13369-024-08949-z (2024).
42. Kumar, N. K., Kumar, V. A., Quamar, D., Reddy, B. S. K., Yogendra, R., & Poojitha, N. (2024, September). Enhanced Alzheimer's disease prediction through advanced imaging: a study of machine learning and deep learning approaches. In 2024 5th International Conference on Smart Electronics and Communication (ICOSEC) (pp. 1177–1182). IEEE. https://doi.org/10.1109/ICOSEC61587.2024.10722116
43. Mubonanyikuzo, V. et al. Detection of Alzheimer disease in neuroimages using vision transformers: Systematic review and meta-analysis. *J. Med. Internet Res.* **27**, e62647. https://doi.org/10.2196/62647 (2025).
44. Biswas, R. & Gini J, R. Multi-class classification of Alzheimer's disease detection from 3D MRI image using ML techniques and its performance analysis. *Multim. Tools Appl.* **83**(11), 33527–33554. https://doi.org/10.1007/s11042-023-16519-y (2024).
45. Raj, A. S., Gunasundari, C., Senthilkumar, S. & Sivamani, S. An optimized hybrid deep learning model to detect Alzheimer disease. *Sci. Rep.* **15**(1), 34081. https://doi.org/10.1038/s41598-025-14169-8 (2025).
46. Alshamlan, H., Alwassel, A., Banafa, A. & Alsaleem, L. Improving Alzheimer's disease prediction with different machine learning approaches and feature selection techniques. *Diagnostics (Basel)* **14**(19), 2237. https://doi.org/10.3390/diagnostics14192237 (2024).
47. Kaya, M. & Cetın-Kaya, Y. A novel deep learning architecture optimization for multiclass classification of Alzheimer's disease level. *IEEE Access* **12**, 46562–46581. https://doi.org/10.1109/ACCESS.2024.3383219 (2024).
48. Breiman, L. Random forests. *Mach. Learn.* **45**(1), 5–32. https://doi.org/10.1023/A:1010933404324 (2001).
49. Rohini, M. & Surendran, D. Weighted hybrid random forest model for significant feature prediction in Alzheimer's disease stages. *Int. J. Comput. Intell. Syst.* **18**(1), 53. https://doi.org/10.1007/s44196-025-00780-0 (2025).
50. Kononenko, I. (1994, April). Estimating attributes: Analysis and extensions of RELIEF. In European conference on machine learning (pp. 171–182). Berlin, Heidelberg: Springer, Berlin and Heidelberg. https://doi.org/10.1007/3-540-57868-4_57
51. Kursa, M. B. & Rudnicki, W. R. Feature selection with the Boruta package. *J. Stat. Softw.* **36**, 1–13. https://doi.org/10.18637/jss.v036.i11 (2010).
52. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**(1), 389–422. https://doi.org/10.1023/A:1012487302797 (2002).
53. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357. https://doi.org/10.1613/jair.953 (2002).
54. Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2623–2631). https://doi.org/10.1145/3292500.3330701
55. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

56. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2 (2020).
57. McKinney, W. Data structures for statistical computing in Python. *Scipy* **445**(1), 51–56. https://doi.org/10.25080/Majora-92bf1922-00a (2010).
58. Lemaıtre, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **18**(17), 1–5. https://doi.org/10.48550/arXiv.1609.06570 (2017).
59. Seabold, S. & Perktold, J. Statsmodels: econometric and statistical modeling with python. *Scipy* **7**(1), 92–96. https://doi.org/10.25080/Majora-92bf1922-011 (2010).
60. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**(3), 90–95. https://doi.org/10.1109/MCSE.2007.55 (2007).
61. Li, M., Liu, H., Li, Y., Wang, Z., Yuan, Y., & Dai, H. (2023, October). Intelligent diagnosis of Alzheimer45s disease based on machine learning. In Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science (pp. 456–462). https://doi.org/10.1145/3644116.3644192
62. Rani, P., Lamba, R., Sachdeva, R. K., Kumar, K. & Iwendi, C. A machine learning model for Alzheimer's disease prediction. *IET Cyber-Phys. Syst.: Theory Appl.* **9**(2), 125–134. https://doi.org/10.1049/cps2.12090 (2024).
63. Hashmi, A. (2026). Early Detection of Alzheimer's Disease Using Explainable Machine Learning on Clinical Biomarkers: A Multi-Class Classification Study Using the Alzheimer's Disease Neuroimaging Initiative (ADNI) Dataset. arXiv preprint arXiv:2606.03995. https://doi.org/10.48550/arXiv.2606.03995
64. Han, L. AD-Diff: Enhancing Alzheimer's disease prediction accuracy through multimodal fusion. *Front. Comput. Neurosci.* https://doi.org/10.3389/fncom.2025.1484540 (2025).
65. Aslan, E. & Özüpak, Y. Comparison of machine learning algorithms for automatic prediction of Alzheimer disease. *J. Chin. Med. Assoc.* **88**(2), 98. https://doi.org/10.1097/JCMA.0000000000001188 (2024).


## Author contributions

RG: Writing – review & editing, Supervision, KT: Methodology, Formal analysis. KKN: Investigation, Formal analysis. HLA: Review & editing, Conceptualization: MVS: Project administration, CSHR: Validation. All authors are involved in editing the final manuscript. The author(s) read and approved the final manuscript.


## Funding

No Funding support has been provided.


## Declarations

## Competing interests

The author declares no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-67252-z.

**Correspondence** and requests for materials should be addressed to H.L.A.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.


**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026