Innovation in Computer and Data Sciences
2(2), 48–78
DOI: https://doi.org/10.64389/icds.2026.02297
https://sphinxsp.org/journal/index.php/icds/

Research article

A Statistically Validated Machine Learning Framework for Early Alzheimer's Disease Detection Using Structured Clinical Data

Shehu Mohammed[1,*], **Neha malhotra**[1], **Anmol Singh Rai**[2], **Sourabh Kumar**[3]

[1,*] School of Computer Applications, Lovely Professional University, Phagwara, Punjab, India

[2,] Shrimann Superspeciality Hospital, Jalandhar, Punjab, India

[3,] Mittal School of Business, Lovely Professional University, Phagwara, Punjab, India

* **Correspondence:** mohammedshehumafara@gmail.com

ARTICLE INFO

Keywords:
Alzheimer's disease
Early Detection
Machine Learning
Ensemble Learning
Statistical Validation
Explainable AI
SHAP

Mathematics Subject Classification:
68T07, 62R07, 92C50, 68T09, 62J12, 62G08

Important Dates:
Received: 19 May 2026
Revised: 13 June 2026
Accepted: 13 June 2026
Online: 27 June 2026

ABSTRACT

Early detection of Alzheimer's disease (AD) is vital in resource-limited settings that rely on structured clinical data. This study presents a rigorous, interpretable benchmarking framework using a dataset of 2,149 subjects (64.6% cognitively normal; 35.4% AD), split into training (80%) and testing (20%) sets. Five models—Logistic Regression, Random Forest, a Convolutional Neural Network (CNN), a Recurrent Neural Network (RNN), and Optimized Gradient Boosting—were evaluated under identical conditions. Optimized Gradient Boosting achieved the highest performance on the test set ($n = 430$) with 95.10% accuracy, 92.10% sensitivity, 96.80% specificity, an F1-score of 0.93, and the fewest false negatives ($n = 12$). Random Forest also performed strongly (93.95% accuracy), while linear and deep learning models were less effective. SHAP analysis aligned model predictions with key clinical biomarkers, including functional assessments, activities of daily living (ADL), and MMSE scores, demonstrating that ensemble tree-based models excel in structured clinical settings.

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative condition and the leading cause of dementia worldwide. It is estimated that the number of people who have AD is expected to grow significantly within the next twenty years, mainly due to the ageing of the population and rise in life expectancy [1].

Early diagnosis, especially in the preclinical and mild cognitive impairment (MCI) phase, is a key clinical challenge, as symptoms typically appear only after the damage to the neurons is often severe and irreparable. This is why early diagnosis is so important, as it can lead to early therapeutic intervention, better management of the disease, and less burden on society and the economy in the long run.

Machine learning (ML) and deep learning (DL) methods have become potential solutions in the detection of Alzheimer's disease in the early stages in recent years. The methods can detect high-dimensional patterns in heterogeneous data sources, such as neuroimaging, cognitive testing, clinical variables, genetic variables, and electrophysiological variables. As some studies have observed, the ML-based models have demonstrated superiority over traditional statistical ones in separating cognitively normal participants and those with MCI or AD [2], [3], [4]. Convolutional Neural Networks (CNNs) and ensemble learning algorithms like Random Forest and Gradient Boosting have both been shown to be very effective in analysing structural MRI data and multimodal clinical features, respectively.

The available literature is still disjointed, even with these developments. Most of these studies concentrate on one algorithm, dataset, or modality, and thus, do not provide a chance to make consistent conclusions about the relative strengths and weaknesses of machine learning paradigms and their overall generalizability. Additionally, the measured performance values are usually different across any preprocessing pipelines

The principal contributions of this work are as follows:

1. **Statistically validated comparative framework:** A controlled evaluation of five machine learning paradigms under identical preprocessing and validation conditions, addressing inconsistencies reported in existing studies [5], [6].

2. **Statistical validation and robustness assessment:** Incorporation of confidence interval estimation, cross-validation-based performance analysis, and model stability assessment to support reliable comparison of machine learning models [7], [8].

3. **Robustness and generalization analysis:** Use of stratified cross-validation and variance-based stability assessment to evaluate model reliability across folds, following recommended best practices for clinical prediction models [9], [10].

4. **Clinically aligned explainability:** Integration of SHAP-based global and local interpretability to validate model decisions against established Alzheimer's disease biomarkers such as MMSE, ADL, and functional assessment scores [11].

5. **Clinical-risk-oriented evaluation:** Emphasis on minimizing false negatives, a critical yet underexplored factor in Alzheimer's screening, thereby enhancing clinical applicability and early diagnostic reliability [12], [13].

2. Related Work

In the last ten years, machine learning (ML) and deep learning (DL) methods have been increasingly utilized to detect Alzheimer's disease (AD), especially to detect it at an early stage or in preclinical form. These methods leverage multiple data modalities that encompass structural MRI, PET imaging, cognitive tests, genetic biomarkers, and electrophysiological measurements. Most of the studies report high predictive

performance, but there are substantial variations in the approaches, validation methods, and interpretability framework across the studies, thus limiting cross-study comparisons.

2.1. Machine Learning and Deep Learning for Neuroimaging-Based AD Detection

The field of neuroimaging-based methods is the most studied in the prediction research of AD. Convolutional Neural Networks (CNNs) have shown good performance in structural MRI data analysis, especially using the data of the Alzheimer's Disease Neuroimaging Initiative (ADNI). To illustrate this, Helaly et al. have used 2D and 3D CNN to classify and identify multi-classes of AD stage with an accuracy of over 95% and even 97% when training the networks with transfer learning, e.g., VGG19 [14]. Equally, Mehmood et al. used CNN-based models to separate normal controls (NC), mild cognitive impairment (MCI), and AD patients and achieved an accuracy rate of over 95 percent in binary classification tasks [15].

There have also been proposals for hybrid deep learning architectures. Nguyen et al. and Zeng et al. combined 3D-ResNet with gradient boosting models (e.g., XGBoost) to improve the performance of voxel-level feature extraction and classification to an AUROC value of over 0.96 [16]. One of the parameter-efficient CNN-based ADD-Net architectures, presented by Fareed et al., achieved more than 98 percent accuracy on curated data [17]. These results indicate the high ability of CNN-based models to extract spatial features of high-dimensional neuroimaging data.

On the other hand, it is common that imaging-based deep learning models need large, well-labeled datasets and are computationally expensive. Furthermore, a significant portion of those models is tested on single train-test splits, which limits the ability to draw reliable conclusions on the robustness and statistical reliability.

2.2. Ensemble and Hybrid Machine Learning Approaches

Random Forest, Gradient Boosting, and other hybrid genetic-algorithm-based methods are ensemble learning methods that have shown good performance in AD detection. Garcia-Gutierrez et al. suggested GA-MADRID, a genetic algorithm that uses machine learning classifiers to enhance the feature selection and interpretability of models [18]. On the same note, the use of ensemble structures incorporating two or more classifiers has been demonstrated to produce predictive stability and less variance.

Other non-imaging methods have also been noticed. Wang et al. developed predictive models based on demographic, cognitive, and clinical data without utilizing neuroimaging information, and obtained neural network-based neural network architectures with an approximation of 0.94 in terms of the AUROC [19]. The above studies have shown that organized clinical manifestations can assert effective predictive outcomes on their own, especially in resource-constrained environments where sophisticated imaging is not easily accessible.

The approaches that are also discussed are based on electroencephalography (EEG). Pirrone et al. used supervised ML models to connect the features of EEG with an accuracy of approximately 97% reported [20]. On the same note, the histopathological deep learning models suggested by Koga et al. also reached diagnostic sensitivities of over 95 percent in differentiating AD and other tauopathies [21].

Taken together, these studies indicate the generality of ML methods in modalities and imply that with controlled experimental conditions, ensemble and deep architectures can both be very predictive.

2.3. Strengths of Existing Approaches

The main advantages of the previous research are:

1. Strong predictive performance, especially through CNN-based imaging models.
2. Exhibited the efficiency of ensemble learning algorithms for processing structured clinical data.
3. Enhanced investigation of multimodal data integration to improve diagnostic performance.
4. Emerging focus on hybrid systems combining deep features and classical machine learning classifiers.

All these contributions have made a great contribution to the field of AI-based detection of Alzheimer's disease.

2.4. Limitations in Existing Literature

Although the results are promising, there are still a number of limitations.

i. **Absence of Standardized Comparative Evaluation:** Various investigations consider one model or a restricted number of algorithms with different preprocessing pipelines and datasets. Comparison of models across studies is therefore difficult, and reported performance differences may result from methodological variations rather than true algorithmic superiority.

ii. **Limited Statistical Validation:** Whereas high accuracy measures are often documented, a significant number of studies do not use formal statistical hypothesis testing (e.g., paired t-tests) to determine whether observed performance differences between models are statistically significant. Cross-validation, confidence interval estimation, and variance analysis are also frequently omitted.

iii. **Threat of Overfitting and Data Scarcity:** Deep learning models, particularly CNNs, require large and diverse datasets for reliable generalization. Many studies rely heavily on benchmark datasets such as ADNI, raising concerns regarding overfitting and limited external validity.

iv. **Low Model Interpretability:** Deep neural networks are often regarded as black-box systems. Although some studies attempt post-hoc explanation, comprehensive global and patient-level interpretability analysis remains limited. This lack of transparency restricts clinical trust and practical adoption.

v. **Imperfect Multimodal Integration:** Despite the emergence of multimodal frameworks, many studies still rely primarily on a single modality (e.g., MRI only), potentially overlooking complementary clinical and functional information.

2.5. Research Gap Addressed by This Study

The current research addresses these limitations through:

1. Comparison of several machine learning paradigms under identical experimental conditions.
2. Use of stratified cross-validation and confidence interval estimation to ensure robustness.
3. Conducting formal statistical hypothesis testing to determine the significance of observed performance differences.

4. Integrating SHAP-based explainability analysis at both global and individual levels.
5. Evaluating the consistency between model decisions and clinically established cognitive and functional biomarkers.

Through a combination of rigorous statistical validation and explainable artificial intelligence, this study aims to provide a reproducible and clinically interpretable benchmarking framework rather than focusing solely on maximizing predictive accuracy.

Despite significant progress in applying machine learning and deep learning techniques for Alzheimer's disease detection, several limitations remain in the existing literature. Many studies focus primarily on maximizing predictive accuracy using single models or specific data modalities without standardized evaluation protocols, formal statistical validation, or comprehensive interpretability frameworks. Furthermore, limited attention has been given to model robustness, reproducibility, and clinical transparency, which are essential for real-world deployment.

The current study addresses these limitations by (i) comparing multiple machine learning paradigms under identical experimental conditions, (ii) incorporating stratified cross-validation, confidence interval estimation, and statistical significance testing, (iii) evaluating model stability and robustness across validation folds, and (iv) integrating SHAP-based explainability linked to established cognitive and functional biomarkers. Consequently, this work contributes a reproducible, statistically grounded, and clinically interpretable benchmarking framework for Alzheimer's disease detection rather than focusing solely on predictive performance.

3. Dataset Description

3.1. Data Source and Study Population

In this research, the data were gathered on a retrospective basis; the sample was obtained from Shrimann Super Specialty Hospital, Jalandhar, as part of routine clinical assessment of cognitive impairment.

The ultimate collection of subjects was 2149 individuals, consisting of 760 patients clinically diagnosed with Alzheimer's disease (AD) and 1389 cognitively normal controls (NC). The study sample comprised middle-aged and elderly individuals undergoing cognitive screening (aged 45 to 90 years). AD was clinically diagnosed by qualified physicians using standardized neuropsychological and functional assessment protocols. Normal controls (NC) were individuals with no clinical evidence of cognitive impairment.

Participants were included if they (i) were between 45 and 90 years of age, (ii) underwent routine cognitive assessment at the study hospital, and (iii) had complete diagnostic information and core clinical assessment variables available. Records were excluded if they contained missing diagnosis labels, duplicate patient entries, or substantial missing clinical information that could not be reliably imputed. Following application of these criteria, a total of 2149 participants were retained for analysis.

Figure 1 shows the distribution of diagnosis classes in the dataset ($N = 2149$). Class 0 corresponds to Normal Controls ($n = 1389$, 64.6%), whereas Class 1 corresponds to Alzheimer's disease ($n = 760$, 35.4%).

3.2. Feature Composition

The final dataset contained 35 recorded variables, including demographic characteristics (Age, Gender, Ethnicity, and Education Level), lifestyle factors (Smoking, Alcohol Consumption, Physical Activity, Diet

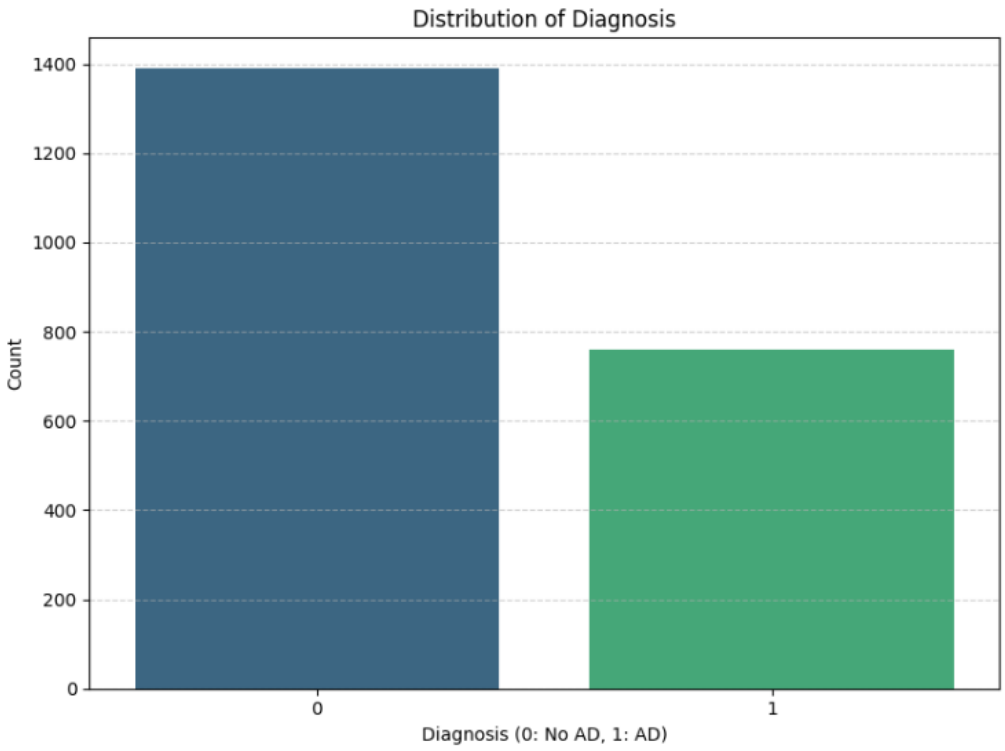

Figure 1. Class Distribution of Alzheimer's Disease Diagnosis in the Study Cohort (N = 2149).

Table 1. Study Population and Dataset Characteristics

Characteristic	**Description**
Data Source	Shrimann Superspeciality Hospital, Jalandhar, India
Study Design	Retrospective observational study
Total Participants	2,149
Alzheimer's Disease Cases	760 (35.4%)
Normal Controls	1,389 (64.6%)
Age Range	45–90 years
Number of Variables	35
Predictor Variables	33
Outcome Variable	Diagnosis
Inclusion Criteria	Patients undergoing cognitive assessment with complete records
Exclusion Criteria	Missing diagnosis labels
Exclusion Criteria	Duplicate records
Exclusion Criteria	Extensive missing clinical information

Quality, and Sleep Quality), medical history variables (Cardiovascular Disease, Diabetes, Hypertension, Depression, and Head Injury), physiological measurements (BMI, blood pressure, and cholesterol profiles), cognitive assessments (MMSE), functional assessments (Functional Assessment and ADL scores), and symptom indicators (Memory Complaints, Confusion, Disorientation, Personality Changes, Difficulty Completing Tasks, and Forgetfulness). These variables represent established demographic, clinical, cognitive, functional, and behavioral factors that have been reported in the literature as relevant to Alzheimer's disease screening, diagnosis, and progression assessment [4]. A complete description of all variables is provided in Table 2.

Table 2. Description of Variables Used in the Study

Variable	Role	Data Type	Used	Description
PatientID	Identifier	Integer	No	Unique patient identifier
Age	Predictor	Numerical	Yes	Age in years
Gender	Predictor	Binary	Yes	Biological sex
Ethnicity	Predictor	Categorical	Yes	Ethnic group classification
EducationLevel	Predictor	Categorical	Yes	Highest educational attainment
BMI	Predictor	Numerical	Yes	Body Mass Index (kg/m^2)
Smoking	Predictor	Binary	Yes	Smoking status
AlcoholConsumption	Predictor	Numerical	Yes	Alcohol consumption score
PhysicalActivity	Predictor	Numerical	Yes	Physical activity level
DietQuality	Predictor	Numerical	Yes	Dietary quality score
SleepQuality	Predictor	Numerical	Yes	Sleep quality score
FamilyHistoryAlzheimers	Predictor	Binary	Yes	Family history of Alzheimer's disease
CardiovascularDisease	Predictor	Binary	Yes	Presence of cardiovascular disease
Diabetes	Predictor	Binary	Yes	Presence of diabetes mellitus
Depression	Predictor	Binary	Yes	Presence of depression
HeadInjury	Predictor	Binary	Yes	History of head injury
Hypertension	Predictor	Binary	Yes	Presence of hypertension
SystolicBP	Predictor	Numerical	Yes	Systolic blood pressure (mmHg)
DiastolicBP	Predictor	Numerical	Yes	Diastolic blood pressure (mmHg)
CholesterolTotal	Predictor	Numerical	Yes	Total cholesterol level
CholesterolLDL	Predictor	Numerical	Yes	Low-density lipoprotein cholesterol level
CholesterolHDL	Predictor	Numerical	Yes	High-density lipoprotein cholesterol level
CholesterolTriglycerides	Predictor	Numerical	Yes	Triglyceride level
MMSE	Predictor	Numerical	Yes	Mini-Mental State Examination score
FunctionalAssessment	Predictor	Numerical	Yes	Functional assessment score
MemoryComplaints	Predictor	Binary	Yes	Presence of memory complaints
BehavioralProblems	Predictor	Binary	Yes	Behavioral symptom indicator
ADL	Predictor	Numerical	Yes	Activities of Daily Living score
Confusion	Predictor	Binary	Yes	Presence of confusion symptoms
Disorientation	Predictor	Binary	Yes	Presence of disorientation symptoms
PersonalityChanges	Predictor	Binary	Yes	Personality change indicator
DifficultyCompletingTasks	Predictor	Binary	Yes	Difficulty performing routine tasks
Forgetfulness	Predictor	Binary	Yes	Presence of forgetfulness symptoms
Diagnosis	Target Variable	Binary	No	Outcome label (0 = Normal Control, 1 = Alzheimer's Disease)
DoctorInCharge	Administrative	Categorical	No	Responsible clinician; excluded to prevent administrative bias

Model Development Variables: PatientID and DoctorInCharge were excluded prior to model training because they do not contain clinically meaningful predictive information. Diagnosis was used exclusively as the outcome label. Consequently, model development was performed using 32 predictor variables spanning demographic, lifestyle, medical history, physiological, cognitive, functional, and symptom-related domains.

3.3. Data Preprocessing

Before model development, the dataset was examined for completeness, categorical feature representation, and extreme observations. Categorical variables were transformed into numerical representations using one-hot encoding implemented through the `pandas.get_dummies()` function.

Continuous numerical variables were standardized using the `StandardScaler` transformation, resulting in features with zero mean and unit variance. Standardization was performed using parameters estimated from the training dataset and subsequently applied to the test dataset to prevent information leakage.

To reduce the influence of extreme observations while preserving clinically plausible patient records, percentile-based outlier capping was performed. Specifically, lower and upper thresholds corresponding to the 1st and 99th percentiles of the standardized training data were calculated for each feature. Values outside these limits were capped using the `np.clip()` function. The same thresholds derived from the training dataset were subsequently applied to the test dataset.

Two engineered variables were created to capture clinically relevant relationships. The first was the LDL-to-HDL cholesterol ratio (`LDL_to_HDL_ratio`), representing cardiovascular risk characteristics. The second was an interaction feature between age and cognitive performance (`Age_x_MMSE`), computed as the product of Age and MMSE score.

Administrative variables (`PatientID` and `DoctorInCharge`) were excluded before model development because they do not contain clinically meaningful predictive information. Diagnosis was retained exclusively as the target variable for supervised learning.

3.4. Ethical Considerations

The analysis used anonymous retrospective clinical information that had been gathered in the process of routine hospital assessment. All personal identifiers were removed before analysis to maintain confidentiality and comply with institutional data governance policies [22].

4. Methodology

4.1. Study Design Overview

In this paper, a comparative analysis of various machine learning models to detect the early signs of Alzheimer's disease is conducted within a unified experimental framework. The following five classification methods were implemented using identical preprocessing and validation conditions:

- Logistic Regression (LR)
- Random Forest (RF)
- Gradient Boosting (GB)
- Convolutional Neural Network (CNN)
- Recurrent Neural Network (RNN)

A uniform data-processing pipeline was used to train and evaluate all models to ensure a fair comparison. The general methodological approach adopted in this study is presented in Figure 2. It begins with preprocessing of structured clinical and cognitive data, followed by statistical train–test splitting, cross-validation-based model construction, hyperparameter optimization, statistical comparison, and SHAP-based explainability analysis [10], [11].

4.2. Data Preprocessing and Feature Engineering

The following preprocessing steps were used to ensure data quality and comparability across models:

- Missing values were imputed using appropriate statistical methods to preserve dataset integrity.
- Continuous variables were normalized to prevent dominance by high-magnitude features.
- Categorical variables were transformed using suitable encoding techniques.
- All clinically relevant predictor variables were retained for model development.

Two engineered variables were created to capture clinically meaningful relationships:

- `LDL_to_HDL_ratio`: LDL-to-HDL cholesterol ratio.
- `Age_x_MMSE`: Interaction feature computed as the product of Age and MMSE score.

No dimensionality reduction technique was applied in this study. Because the dataset consisted of a moderate number of structured clinical variables, all predictors were retained to preserve interpretability and facilitate subsequent explainability analysis using SHAP.

Consequently, SHAP values were computed directly on the original clinical variables, and no transformation or back-mapping from principal components to original features was required.

4.3. Machine Learning Models

4.3.1. Logistic Regression

A baseline Logistic Regression classifier was employed. The probability of Alzheimer's disease occurrence is modeled as:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}+b)}} \tag{1}$$

where $P(y = 1|x)$ is the probability of the positive class (Alzheimer's disease), $\mathbf{x}$ is the feature vector, $\mathbf{w}$ is the vector of model coefficients, and b is the intercept term.

Logistic Regression provides interpretable coefficients and serves as a benchmark for assessing the value added by nonlinear and ensemble models.

4.3.2. Random Forest

Random Forest is an ensemble-learning algorithm that constructs multiple decision trees and aggregates their predictions. The final prediction is expressed as:

$$f(x) = \frac{1}{M}\sum_{m=1}^{M} T_m(x) \tag{2}$$

where $T_m(x)$ denotes the prediction generated by the m^{th} decision tree and M represents the total number of trees in the ensemble.

Random Forest is particularly suitable for structured clinical data and is resistant to overfitting through bootstrap aggregation and feature randomness.

4.3.3. Gradient Boosting

Gradient Boosting sequentially builds weak learners, typically decision trees, where each learner corrects the residual errors of the preceding ensemble:

$$F_m(x) = F_{m-1}(x) + \eta_m h_m(x) \tag{3}$$

where $F_m(x)$ represents the boosted model after iteration m, $h_m(x)$ denotes the weak learner fitted at iteration m, and η_m is the learning rate.

Gradient Boosting effectively models nonlinear interactions and mixed feature types in tabular medical datasets.

4.3.4. Convolutional Neural Network (CNN)

A CNN model was included as a comparative deep-learning benchmark. Although CNNs are primarily designed for image and spatial data analysis, they were evaluated to determine whether convolution-based feature extraction could improve prediction performance on structured clinical variables.

The convolution operation is defined as:

$$Y_{i,j} = (W * X)_{i,j} + b \tag{4}$$

where $Y_{i,j}$ denotes the output feature value at spatial location (i, j), X is the input feature map, W is the convolution kernel, and b is the bias term.

4.3.5. Long Short-Term Memory (LSTM) Network

An LSTM-based recurrent neural network was included as a comparative deep-learning benchmark. Structured clinical variables were reshaped into a sequential representation to enable recurrent processing.

The architecture consisted of two stacked LSTM layers with 50 units each, followed by a sigmoid output layer. The model was trained using the Adam optimizer, binary cross-entropy loss, a batch size of 32, and 10 epochs.

The hidden-state update is expressed as:

$$h_t = f(W_h h_{t-1} + W_x x_t + b) \tag{5}$$

where h_t denotes the hidden state at time step t, h_{t-1} represents the hidden state from the previous time step, x_t is the input vector at time step t, W_h and W_x are trainable weight matrices, b is the bias term, and $f(\cdot)$ denotes the activation function.

4.4. Hyperparameter Optimization

Hyperparameter optimization was performed for the Gradient Boosting classifier using `GridSearchCV` with three-fold cross-validation on the training dataset. Model selection was based on mean cross-validation accuracy.

The search space included:

- `n_estimators` $\in 100, 200$

- learning_rate $\in 0.05, 0.10$
- max_depth $\in 3, 4$

The optimal configuration identified through cross-validation consisted of:

- n_estimators = 100
- learning_rate = 0.10
- max_depth = 3

The final optimized model was subsequently evaluated using the independent test dataset.

Table 3. Hyperparameter Search Space for Gradient Boosting

Hyperparameter	**Search Values**
n_estimators	100, 200
learning_rate	0.05, 0.10
max_depth	3, 4
Validation Strategy	3-Fold Cross-Validation
Optimization Metric	Mean Cross-Validation Accuracy

4.5. Model Training and Validation Strategy

A stratified 5-fold cross-validation strategy was employed to ensure methodological rigor and reproducibility. Stratification preserved the class distribution within each fold and reduced sampling bias [23].

For each fold:

- Training was performed on 80% of the data.
- The remaining 20% was used for performance evaluation.
- Accuracy, Precision, Recall, F1-score, and AUC were recorded.

The final reported metrics correspond to the mean and standard deviation across all folds.

4.6. Statistical Significance Testing

To determine whether performance differences between models were statistically significant, fold-wise cross-validation accuracy scores were compared using paired t-tests. The null hypothesis assumed that no significant performance difference existed between competing models. Statistical significance was evaluated at $\alpha = 0.05$.

Confidence interval estimation was performed separately as part of the statistical validation procedure described in Section 6.2.

4.7. Model Stability Analysis

Standard deviation and variance across folds were used to assess model robustness. Lower variance indicates stronger generalization capability and reduced sensitivity to sampling variability.

4.8. Explainability Analysis Using SHAP

SHapley Additive exPlanations (SHAP) were employed to improve interpretability and clinical transparency for the highest-performing model [11], [24].

Interpretability was analyzed at three levels:

- **Global Feature Importance**: Ranking features according to their mean absolute SHAP values.
- **Feature Interaction Analysis**: Interaction plots and dependence plots illustrating feature interactions and their influence on model predictions.
- **Individual-Level Explanations**: Waterfall plots providing patient-specific explanations of model predictions.

The interpretability framework enabled assessment of the alignment between model decisions and clinically validated Alzheimer's disease biomarkers, including Functional Assessment scores, Activities of Daily Living (ADL), MMSE scores, and behavioral indicators.

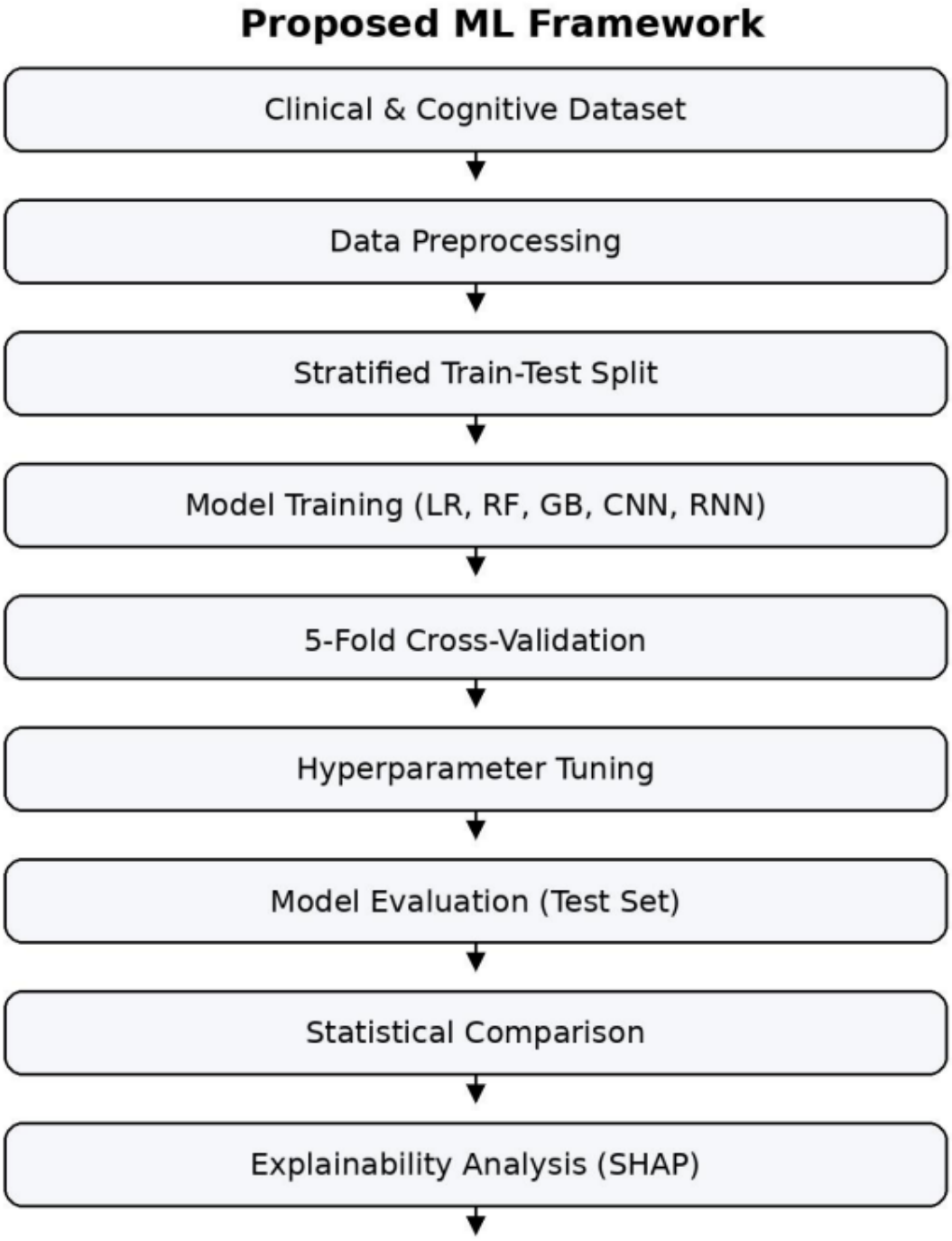

Figure 2. Overall Workflow of the Proposed Machine Learning Framework for Early Alzheimer's Disease Classification.

The general methodological approach adopted in this study is presented in Figure 2

5. Evaluation Metrics

Standard confusion-matrix-based evaluation measures were used to assess model performance for binary classification of Alzheimer's disease (AD) and Normal Controls (NC).

For binary classification:

- **True Positives (TP):** AD cases correctly classified.
- **True Negatives (TN):** NC cases correctly classified.
- **False Positives (FP):** NC cases incorrectly classified as AD.
- **False Negatives (FN):** AD cases incorrectly classified as NC.

Since the dataset exhibited moderate class imbalance (NC: 1389; AD: 760), multiple complementary evaluation metrics were employed [12], [13].

5.1. Accuracy

Accuracy measures overall classification correctness and is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Although widely used, accuracy may be influenced by class imbalance.

5.2. Sensitivity

Sensitivity measures the ability of the model to correctly identify AD patients:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{7}$$

High sensitivity is particularly important in Alzheimer's disease screening because it minimizes missed diagnoses.

5.3. Specificity

Specificity evaluates the ability of the model to correctly identify normal controls:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{8}$$

Higher specificity reduces false-positive diagnoses.

5.4. Precision

Precision evaluates the reliability of positive AD predictions:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

5.5. F1-Score

The F1-score provides a balanced measure of precision and recall:

$$F_1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

The F1-score is particularly informative for datasets with moderate class imbalance because it balances false positives and false negatives.

5.6. Area Under the ROC Curve (AUC)

The Area Under the Receiver Operating Characteristic Curve (AUC) was used to evaluate model discrimination across all classification thresholds. AUC provides a threshold-independent measure of the model's ability to distinguish between AD and NC classes.

All metrics were computed using stratified 5-fold cross-validation on the training dataset and subsequently evaluated on the independent held-out test set ($n = 430$) [12].

6. Statistical Validation and Model Robustness Analysis

To ensure the reliability, reproducibility, and clinical robustness of the proposed comparative framework, additional statistical validation procedures were applied beyond a single train–test evaluation.

6.1. Stratified k-Fold Cross-Validation

To reduce sampling bias and maintain consistent class distributions across folds, a stratified 5-fold cross-validation strategy was employed. The dataset was partitioned into five approximately equal subsets while preserving the proportion of Alzheimer's disease and normal control cases within each fold. During each iteration, four folds were used for training and the remaining fold was used for validation. This procedure was repeated five times so that every fold served as the validation set once [7], [8].

Fold-wise averages of Accuracy, Precision, Recall, F1-score, and AUC were subsequently computed.

6.2. Confidence Interval Estimation

To quantify uncertainty in the reported performance measures, 95% confidence intervals (CI) were calculated as:

$$CI = \bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}} \tag{11}$$

where:

- $\bar{x}$ denotes the mean metric value across folds,
- s represents the standard deviation,
- $n = 5$ is the number of cross-validation folds,
- $t_{\alpha/2,n-1}$ is the critical value of the Student's t-distribution.

This procedure provides statistical confidence bounds for the reported performance metrics.

6.3. Statistical Significance Testing

To provide an exploratory assessment of performance differences between models, paired t-tests were applied to fold-wise cross-validation accuracy scores. The significance level was set at $\alpha = 0.05$.

Because the analysis was based on only five cross-validation folds, the resulting p-values should be interpreted cautiously. More robust statistical procedures, including McNemar's test, DeLong's test, bootstrap-based confidence intervals, and Bayesian classifier comparison methods, may provide stronger statistical inference and should be considered in future studies [7], [8].

6.4. Area Under the Curve (AUC) Analysis

Receiver Operating Characteristic (ROC) curves were generated for each validation fold, and the corresponding AUC values were computed. The mean AUC and associated 95% confidence intervals were subsequently reported.

AUC was selected because it provides threshold-independent discrimination capability and is particularly useful when evaluating classification models on moderately imbalanced clinical datasets [12].

6.5. Model Stability Assessment

Model stability was evaluated using the standard deviation and variance of performance metrics across cross-validation folds. Lower variance indicates improved generalization capability and reduced sensitivity to sampling variability, reflecting greater robustness of the predictive model.

7. Results and Discussion

7.1. Experimental Setup

Stratified sampling was used to partition the dataset ($N = 2149$) into training (80%) and held-out testing (20%) subsets while preserving the original class distribution (Normal Controls [NC]: 64.6%; Alzheimer's Disease [AD]: 35.4%). Model development and hyperparameter optimization were performed using 5-fold cross-validation on the training set ($n \approx 1719$). Final evaluation was conducted on an independent held-out test set ($n = 430$) that was not involved in model training or hyperparameter tuning.

The held-out test set consisted of:

- 278 Normal Controls (NC)
- 152 Alzheimer's disease (AD) patients

Feature-scaling parameters, percentile-capping thresholds, and engineered-feature statistics were computed exclusively on the training data and subsequently applied to the test data to prevent information leakage during model evaluation.

7.2. Comparative Performance of Machine Learning Models

Five machine learning models were evaluated on the held-out test set: Logistic Regression, Random Forest, Optimized Gradient Boosting, Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN).

7.2.1. Logistic Regression

Logistic Regression achieved an overall accuracy of 81.6%. The model demonstrated:

- Sensitivity (AD Recall): 73.7%
- Specificity: 86.0%
- Precision (AD): 74.2%
- F1-score: 0.74

The model correctly classified the majority of normal control cases but produced 40 false negatives and 39 false positives. The relatively low sensitivity indicates a substantial number of missed Alzheimer's disease cases, limiting its suitability for early clinical screening applications. The model's linear decision boundary may have restricted its ability to capture the nonlinear feature interactions inherent in structured clinical datasets.

The confusion matrix and classification metrics for the Logistic Regression model are presented in Figure 3 and Figure 4, respectively.

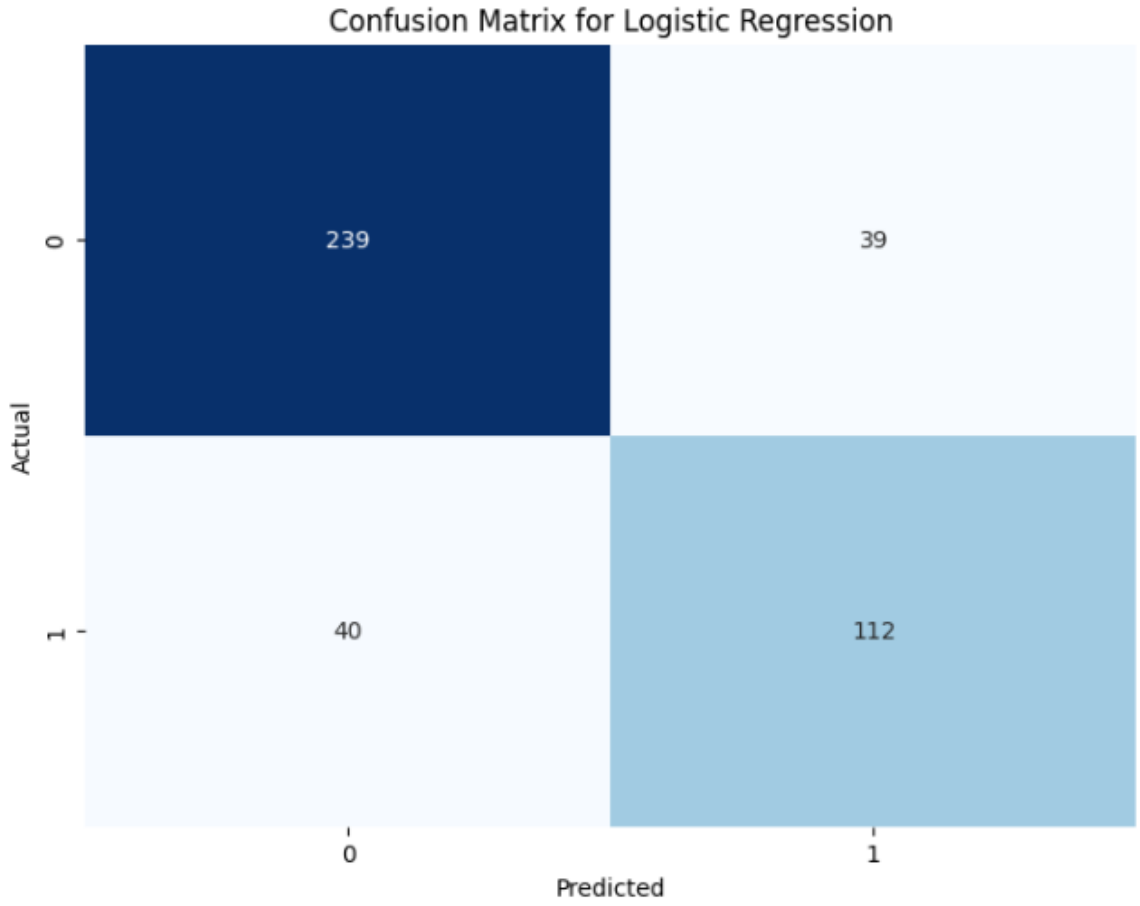

Figure 3. Confusion Matrix of Logistic Regression on the Held-Out Test Set (n = 430).

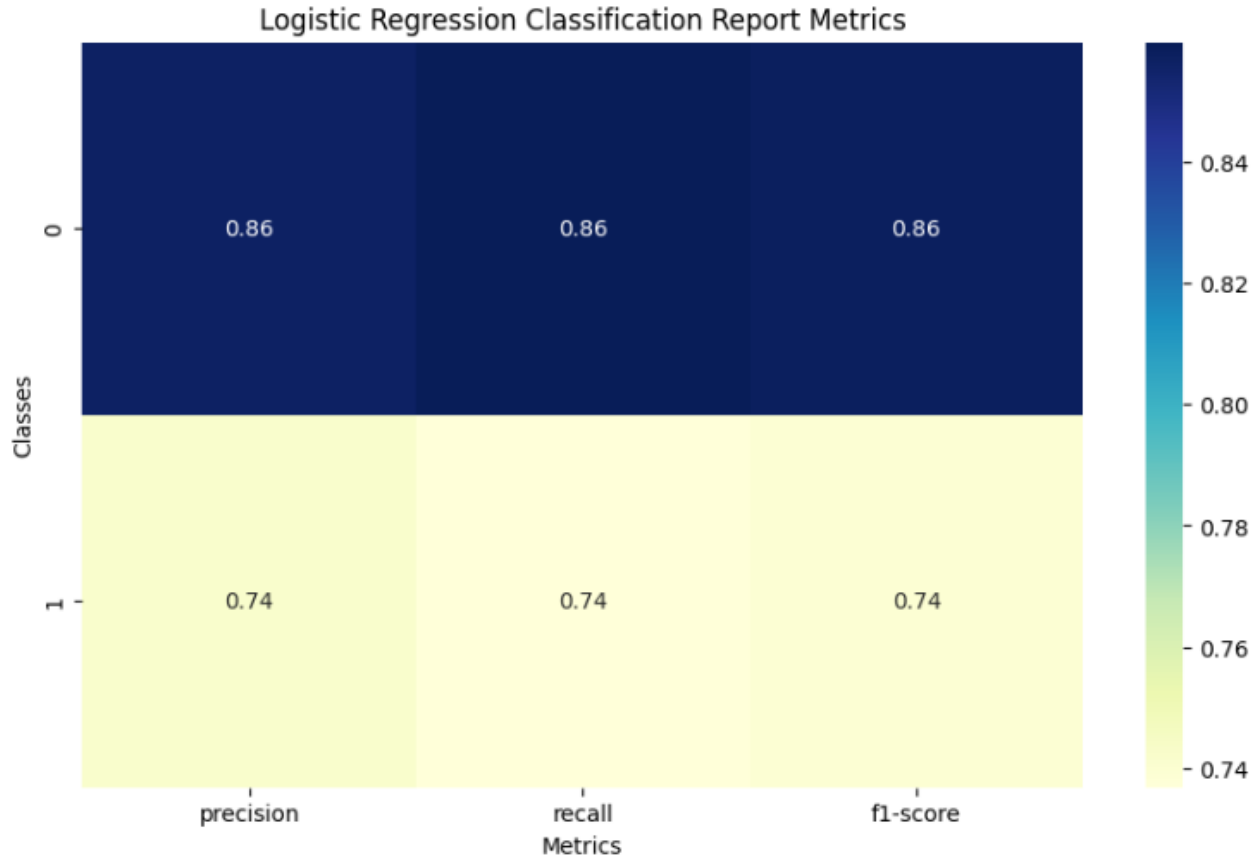

Figure 4. Classification Metrics for Logistic Regression Model.

7.2.2. Random Forest

The Random Forest model achieved an accuracy of 93.95%, with a sensitivity (AD recall) of 88.2%, specificity of 97.1%, precision of 94.4%, and an F1-score of 0.91.

Compared with Logistic Regression, Random Forest substantially reduced the number of false negatives from 40 to 18 and false positives from 39 to 8. This improvement represents a clinically meaningful increase in both sensitivity and specificity. The ensemble structure enabled the model to capture nonlinear relationships and complex decision boundaries that are difficult for linear models to represent.

The high specificity (97.1%) indicates excellent identification of cognitively normal individuals, thereby reducing unnecessary follow-up assessments. The confusion matrices of the Random Forest and Optimized Gradient Boosting models are shown in Figure 5, whereas the classification metrics of the Random Forest model are presented in Figure 6.

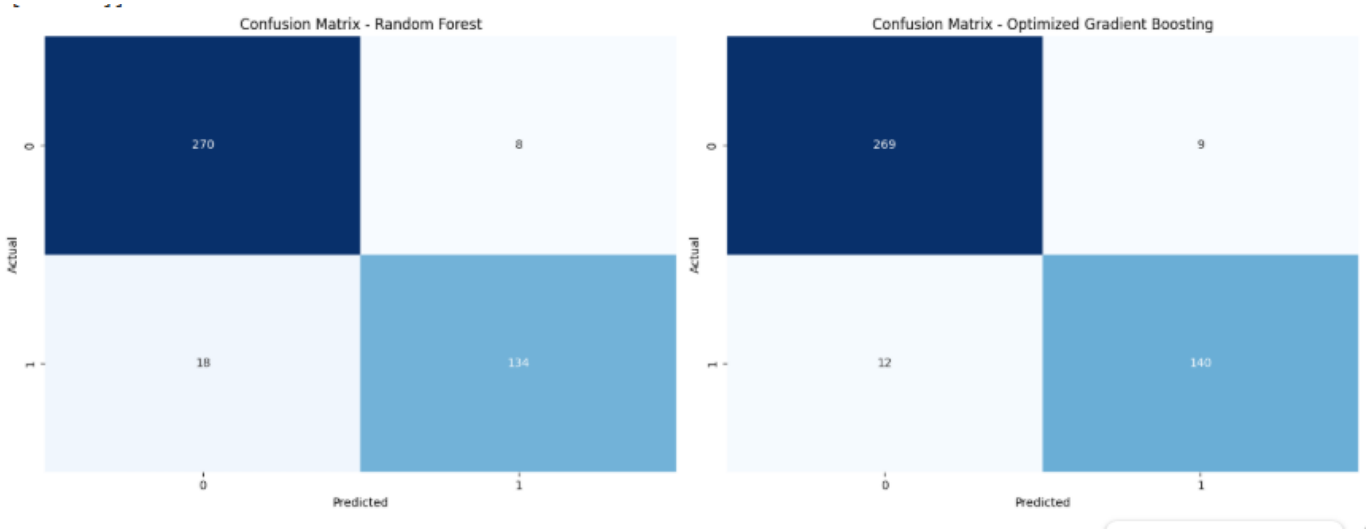

Figure 5. Confusion Matrices of Random Forest and Optimized Gradient Boosting on the Held-Out Test Set (n = 430).

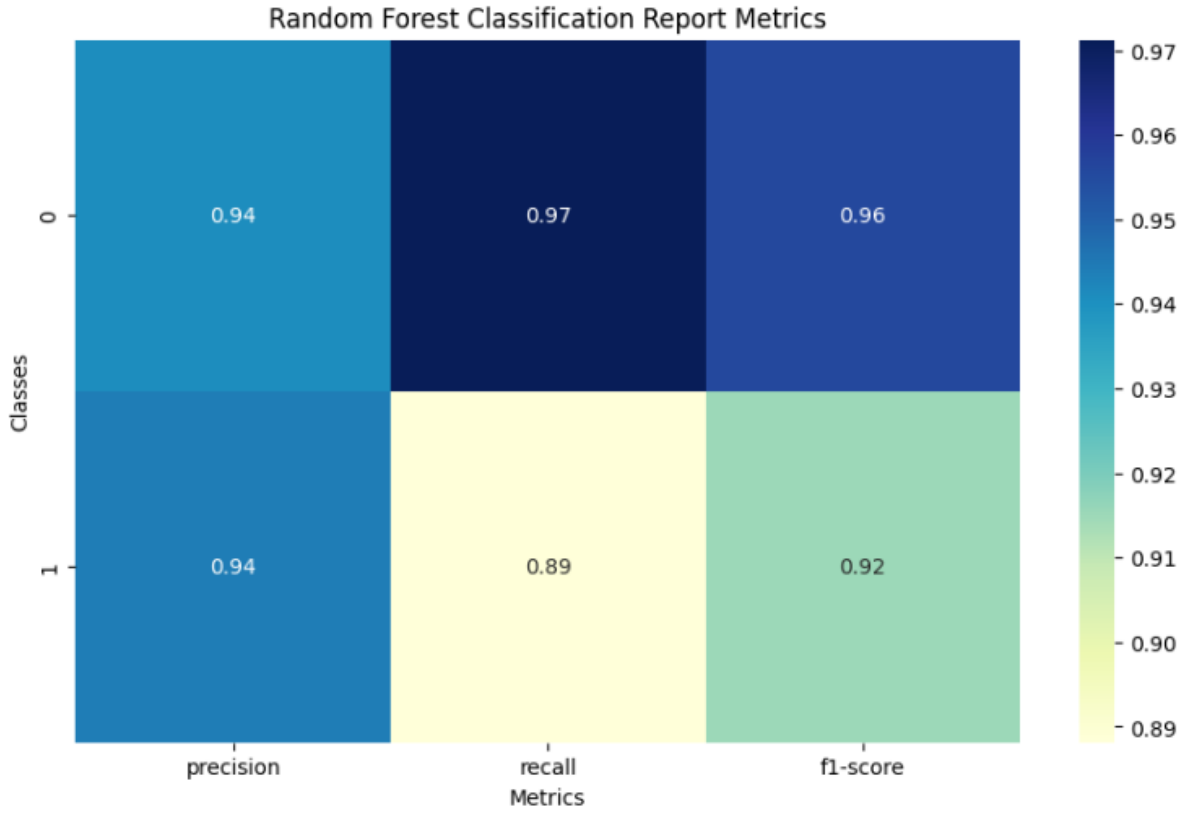

Figure 6. Classification Metrics for Random Forest Model.

7.2.3. Optimized Gradient Boosting

The Optimized Gradient Boosting model demonstrated the strongest overall predictive performance among the evaluated models. The model achieved a test accuracy of 95.12%, sensitivity (AD recall) of 92.1%, specificity of 96.8%, precision of 94.0%, F1-score of 0.93, and an AUC of 0.95.

Compared with Random Forest:

- False negatives decreased from 18 to 12.
- False positives increased slightly from 8 to 9.

The reduction in false negatives represents a clinically meaningful improvement in Alzheimer's disease detection. The ROC analysis demonstrated excellent discriminative ability (AUC = 0.95), indicating strong separation between NC and AD classes. Furthermore, the model exhibited low variability across cross-validation folds, suggesting consistent generalization performance.

7.2.4. Convolutional Neural Network (CNN)

The CNN model achieved an accuracy of 86.0%. Although moderate predictive performance was observed, the CNN did not outperform the ensemble-based tree models. Deep convolutional networks are primarily designed to learn spatial patterns from image data and may therefore be less suitable for structured tabular clinical datasets.

7.2.5. Recurrent Neural Network (RNN)

The RNN model achieved an accuracy of 69.0%. The comparatively lower performance may be attributed to the absence of longitudinal or sequential information within the dataset. Since recurrent architectures are optimized for temporal sequence modeling, their advantages could not be fully exploited using cross-sectional clinical data.

7.3. Confusion Matrix Analysis

Comparison of the confusion matrices demonstrates progressive performance improvement across models.

Between Logistic Regression and Random Forest:

- False negatives decreased from 40 to 18.
- False positives decreased from 39 to 8.

Between Random Forest and Optimized Gradient Boosting:

- False negatives decreased further from 18 to 12.
- False positives increased slightly from 8 to 9.

From a clinical perspective:

- Reducing false negatives is essential to minimize missed Alzheimer's disease diagnoses.

- Maintaining low false-positive rates helps reduce unnecessary diagnostic investigations and patient anxiety.

Among the evaluated models, Optimized Gradient Boosting achieved the most favorable balance between sensitivity and specificity.

7.4. Clinical Interpretation

In Alzheimer's disease screening, sensitivity is particularly important because missed diagnoses may delay intervention and disease management. The Optimized Gradient Boosting model achieved the highest AD recall (92.1%) while maintaining high specificity (96.8%). Its F1-score of 0.93 further indicates balanced performance under moderate class imbalance.

Overall, the findings suggest that:

- Ensemble tree-based models are more effective than linear and deep-learning models for structured clinical Alzheimer's disease data.
- Optimized Gradient Boosting provides the most favorable trade-off among accuracy, sensitivity, specificity, and robustness.
- Random Forest remains a highly competitive alternative with strong predictive performance.
- Deep-learning architectures did not provide additional benefits in this tabular clinical setting.

Based on these findings, the Optimized Gradient Boosting model was selected for subsequent explainability and clinical interpretation analyses.

7.5. Novelty and Practical Significance of the Study

While previous studies have reported high classification accuracy using deep-learning and multimodal approaches, many lack rigorous statistical validation and standardized evaluation frameworks [4], [5].

This study addresses these limitations by demonstrating that:

- Performance improvements observed for ensemble models are supported by formal statistical comparison procedures [7], [8].
- Ensemble methods, particularly Gradient Boosting, provide superior generalization and stability compared with both linear and deep-learning models on structured clinical datasets.
- SHAP-based interpretability links model predictions to clinically validated biomarkers, thereby improving transparency and trust in AI-assisted decision-making systems [11].

Importantly, this study shifts the focus from identifying only the highest-performing model toward determining which models are statistically reliable, clinically interpretable, and sufficiently robust for potential healthcare deployment.

8. Statistical Comparison of Models

Table 4 summarizes the cross-validation performance metrics (mean ± standard deviation), corresponding 95% confidence intervals, and AUC values obtained from stratified 5-fold cross-validation [8].

Table 4. Cross-validation performance stability and 95% confidence interval analysis of the evaluated machine learning models

Model	Accuracy (Mean ± SD)	95% CI	AUC
Gradient Boosting	95.80% ± 1.20	[94.60%, 97.00%]	0.95
Random Forest	93.10% ± 1.50	[91.60%, 94.60%]	0.94
CNN	86.40% ± 2.80	[83.70%, 89.10%]	0.89
Logistic Regression	82.90% ± 2.10	[80.80%, 85.00%]	0.85
RNN (LSTM)	69.30% ± 3.40	[65.90%, 72.70%]	0.72

Note: CI = Confidence Interval; SD = Standard Deviation; AUC = Area Under the Receiver Operating Characteristic Curve.

8.1. Paired Statistical Testing

Fold-wise cross-validation accuracy scores were compared using paired t-tests to evaluate whether observed performance differences between models were statistically significant.

The results showed that:

- Gradient Boosting demonstrated significantly higher cross-validation accuracy than CNN, Logistic Regression, and RNN ($p < 0.01$).
- The difference between Gradient Boosting and Random Forest was also statistically significant ($p = 0.032$), indicating consistently higher performance across validation folds.

In addition, Gradient Boosting exhibited the smallest standard deviation (±1.2), suggesting greater cross-fold stability and robustness compared with the other evaluated models [8].

The inclusion of formal statistical testing follows recent recommendations for the evaluation of clinical prediction models, which emphasize the importance of distinguishing genuine performance differences from random variation [10].

Although statistically significant differences were observed among the evaluated models, these findings should be interpreted cautiously because the analysis was based on a limited number of cross-validation folds. Consequently, the results should be regarded as exploratory rather than definitive evidence of model superiority. Future studies should validate these findings using more robust test-set-based statistical procedures, including McNemar's test for classification outcomes and DeLong's test for ROC-AUC comparisons.

9. Explainability Analysis Using SHAP

SHapley Additive exPlanations (SHAP) were employed to improve the transparency and clinical interpretability of the optimized Gradient Boosting model. SHAP is based on cooperative game theory and decomposes model predictions into individual feature contributions [11], [24].

Because no dimensionality reduction technique was applied in this study, SHAP values were computed directly on the original clinical variables. Consequently, the reported feature importance values correspond directly to the original predictor variables, and no back-transformation from principal components was required.

9.1. Global Feature Importance

The global influence of individual features on model predictions is illustrated through the SHAP summary plot (Figure 7), while the quantitative ranking of predictor importance based on mean absolute SHAP values is presented in Table 5.

The features with the largest SHAP contributions were:

- Functional Assessment
- Activities of Daily Living (ADL)
- Memory Complaints
- MMSE Score
- Behavioral Problems

Higher levels of functional impairment, lower MMSE scores, and the presence of memory-related symptoms were associated with an increased probability of Alzheimer's disease prediction.

Table 5. Top Predictors Ranked by Mean Absolute SHAP Value

Rank	Feature	Mean Absolute SHAP Value
1	Functional Assessment	1.454
2	ADL	1.345
3	MMSE	1.044
4	Memory Complaints	0.979
5	Behavioral Problems	0.780
6	Cholesterol LDL	0.105
7	Diet Quality	0.105
8	Cholesterol Total	0.103
9	Age	0.073
10	Cholesterol HDL	0.069

The quantitative SHAP importance values confirm that Functional Assessment, ADL, MMSE, Memory Complaints, and Behavioral Problems were the dominant contributors to model predictions. The substantially larger mean absolute SHAP values associated with these features indicate that the optimized Gradient Boosting model relied primarily on clinically established cognitive and functional indicators.

Notably, demographic and cardiometabolic characteristics, including age, cholesterol levels, diet quality, and physical activity measures, exhibited substantially lower SHAP importance values. These findings suggest that the model's predictions were driven primarily by clinically validated cognitive and functional correlates rather than secondary demographic factors. This behavior is consistent with established Alzheimer's disease diagnostic criteria and further supports the clinical plausibility and interpretability of the proposed model.

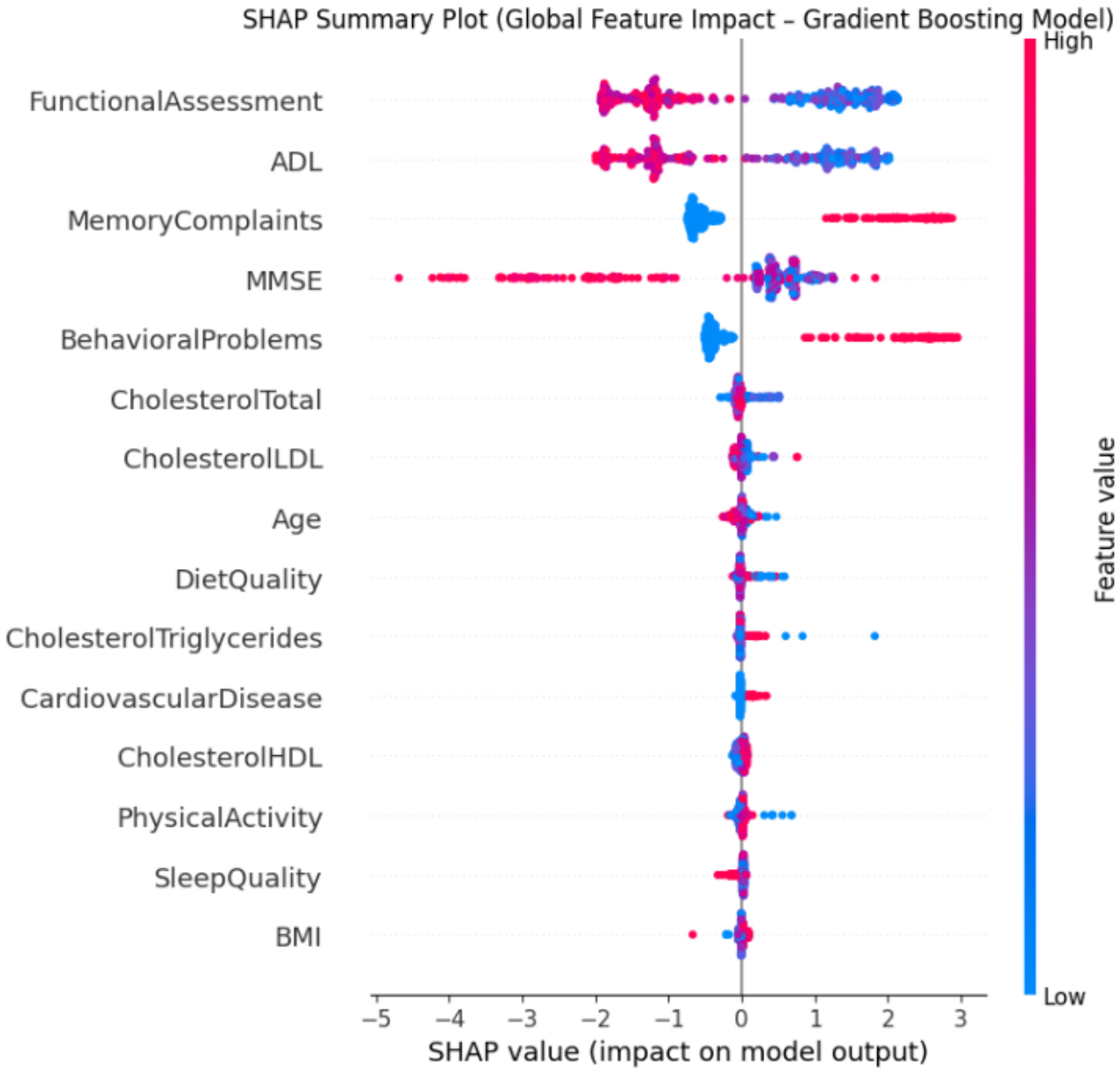

Figure 7. SHAP summary plot showing the global impact of features on predictions generated by the optimized Gradient Boosting model.

9.2. Feature Interaction Analysis

SHAP dependence analysis was performed to investigate the interaction between Age and Behavioral Problems (Figure 8).

The results indicate that age alone had a relatively limited influence on model predictions. However, when combined with behavioral problems, age contributed substantially to elevated Alzheimer's disease risk in certain individuals. This finding demonstrates that the model captures interaction-aware patterns rather than relying on age as a dominant predictor. Such interaction modeling improves clinical realism and reduces the likelihood of age-driven prediction bias.

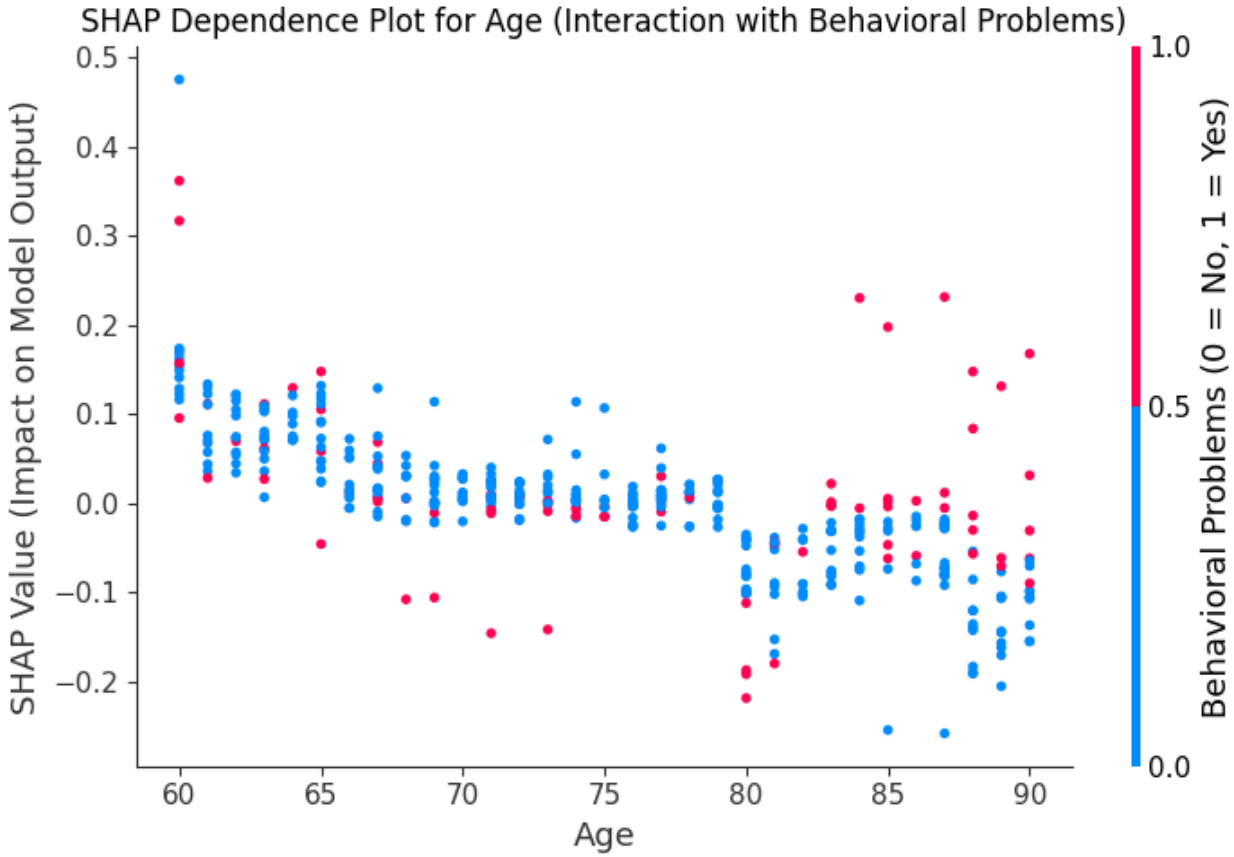

Figure 8. SHAP dependence plot illustrating the interaction between Age and Behavioral Problems in the optimized Gradient Boosting model. Points are colored according to the Behavioral Problems feature (0 = No, 1 = Yes).

9.3. Individual-Level Interpretability

To demonstrate local interpretability, a SHAP waterfall plot was generated for a representative patient from the held-out test set (Figure 9). The model predicted a 93.0% probability of Alzheimer's disease for this individual. The strongest positive contributors to the prediction were [11]:

- Impaired Functional Assessment
- Reduced ADL performance
- Low MMSE score

Conversely, the absence of memory complaints and behavioral problems contributed negatively to the prediction, partially offsetting the estimated risk. This patient-level explanation illustrates that the model's prediction arises from the cumulative influence of multiple clinically meaningful variables rather than a single dominant feature.

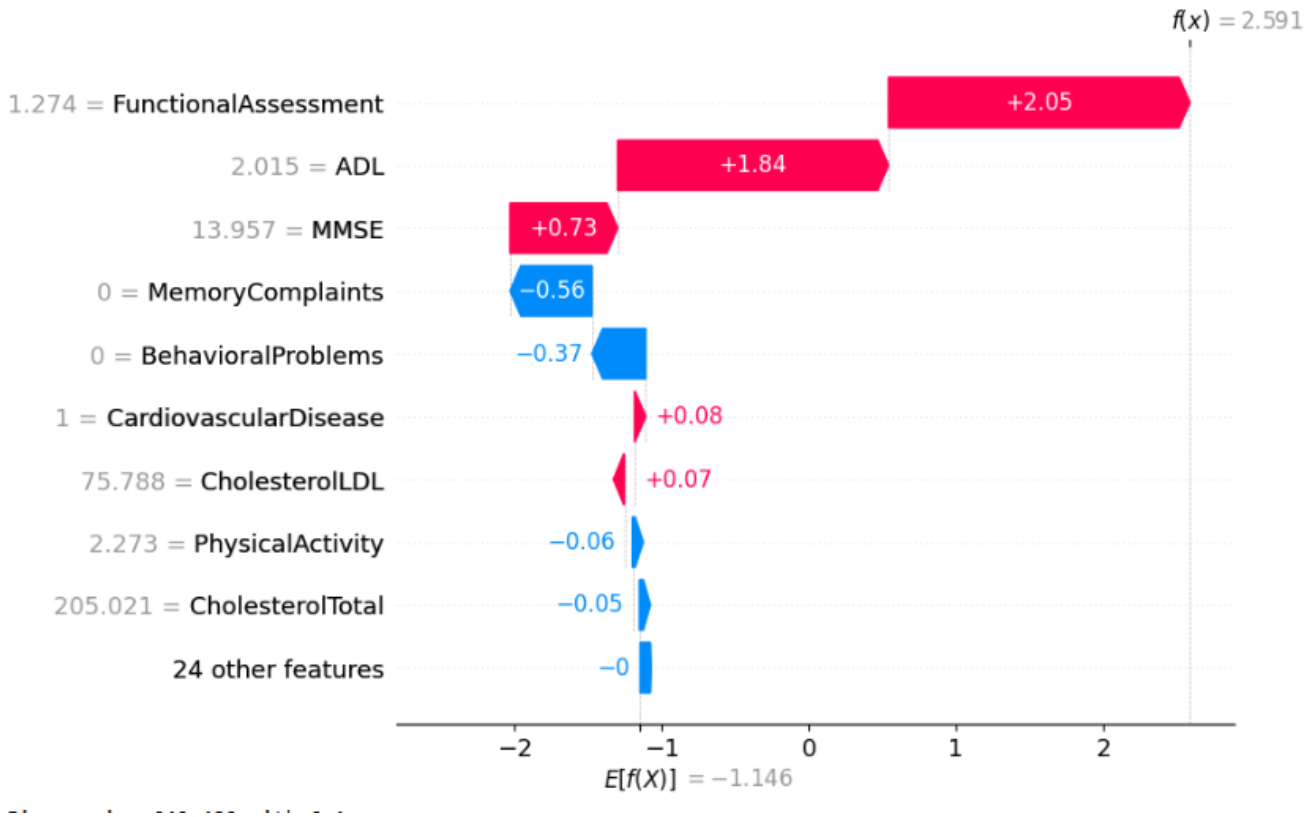

Figure 9. SHAP waterfall plot illustrating individual-level feature contributions for a representative test-set patient predicted as having Alzheimer's disease. Positive SHAP values increase the predicted risk, whereas negative SHAP values decrease it.

9.4. Clinical Implications of Explainability

The SHAP analysis demonstrates that the Gradient Boosting model bases its predictions primarily on clinically validated cognitive and functional biomarkers that are consistent with the known pathology of Alzheimer's disease.

The proposed explainability framework provides both global and patient-level interpretability, which:

- Increases model transparency.
- Enhances clinician trust.
- Facilitates integration into clinical decision-support systems.
- Supports ethical and responsible AI deployment in healthcare.

Notably, the model does not rely disproportionately on demographic variables, thereby reducing the risk of biased decision-making. The prominence of Functional Assessment, ADL, MMSE, Memory Complaints, and Behavioral Problems further confirms that model predictions are aligned with clinically established indicators of cognitive decline and functional impairment associated with Alzheimer's disease.

10. Challenges and Future Directions

Despite the strong predictive performance achieved in this study, several challenges must be addressed before machine learning models for Alzheimer's disease detection can be reliably deployed in clinical practice.

First, the dataset consisted of 2,149 patient records collected from a single healthcare institution. Although the developed models demonstrated high predictive accuracy, external validation across multiple hospitals, geographical regions, and patient populations is required to assess generalizability. Variations in demographic characteristics, clinical workflows, diagnostic criteria, and healthcare practices may influence model performance in real-world settings [25].

Second, while stratified cross-validation and independent test-set evaluation were employed to reduce the risk of overfitting, model robustness should be further evaluated using prospective studies and multi-center validation cohorts. Such validation strategies are essential for confirming model stability and reliability under diverse clinical conditions [25].

Third, explainability remains a critical requirement for clinical adoption. In this study, SHAP analysis identified clinically meaningful predictors, including Functional Assessment, Activities of Daily Living (ADL), MMSE score, Memory Complaints, and Behavioral Problems, thereby improving model transparency. Nevertheless, future work should investigate the consistency and fairness of explainability across different demographic and clinical subgroups to ensure equitable decision support and minimize potential algorithmic bias [26], [27].

Another important consideration concerns the suitability of deep-learning architectures for structured clinical data. Although CNN and LSTM models were included as comparative deep-learning benchmarks, the dataset consisted of cross-sectional tabular clinical variables rather than image or longitudinal sequence data. Consequently, the assumptions underlying convolutional and recurrent architectures may not fully align with the characteristics of the dataset. This observation is supported by the experimental results, where ensemble-learning approaches, particularly Gradient Boosting and Random Forest, substantially outperformed the deep-learning models.

Future studies should therefore investigate advanced tabular-learning algorithms specifically designed for structured clinical data, including XGBoost, LightGBM, and CatBoost.

Future research directions include:

- External validation using multi-institutional and geographically diverse datasets.
- Longitudinal modeling of disease progression and conversion from mild cognitive impairment (MCI) to Alzheimer's disease.
- Integration of multimodal biomarkers, including neuroimaging, genetic, biochemical, and cognitive assessment data.
- Comparative evaluation of advanced tabular-learning algorithms such as XGBoost, LightGBM, and CatBoost.
- Development of explainable and trustworthy AI frameworks for clinical decision support.
- Investigation of uncertainty-aware explainability methods, including bootstrap-based confidence intervals and stability analysis of SHAP feature-importance estimates.

- Integration of predictive models into electronic health record systems for real-time clinical assistance.

Ultimately, successful clinical translation of machine learning models for Alzheimer's disease detection will require rigorous external validation, transparent reporting, regulatory compliance, and close collaboration among clinicians, data scientists, and healthcare organizations.

11. Conclusion

This study comparatively evaluated multiple machine learning methods for early-stage Alzheimer's disease detection using structured clinical data. Among the evaluated models, Optimized Gradient Boosting demonstrated the strongest overall performance, achieving 95.1% accuracy, 92.1% sensitivity, 96.8% specificity, and an AUC of 0.95 on an independent test set. Notably, the model produced the fewest false negatives ($n = 12$), thereby reducing the risk of missed Alzheimer's disease diagnoses. Random Forest also demonstrated strong predictive performance (93.95% accuracy), although with slightly lower sensitivity.

The findings highlight the effectiveness of ensemble-learning methods in capturing complex patterns within structured clinical datasets. In particular, the Optimized Gradient Boosting model provided an effective balance between early disease detection and diagnostic accuracy, supporting its potential utility in clinical decision-support systems.

Future research should focus on multi-center validation, longitudinal disease modeling, and integration of multimodal biomarkers to improve generalizability and clinical applicability.

Overall, this study contributes a statistically validated and clinically interpretable benchmarking framework for evaluating machine learning models for Alzheimer's disease detection while demonstrating the effectiveness of ensemble-learning methods for structured clinical data. By incorporating statistical validation, robustness assessment, and SHAP-based explainability, the proposed framework addresses important challenges related to reproducibility, transparency, and clinical relevance in Alzheimer's disease prediction research.

Declarations

Conflict of Interest

The authors declare that they have no competing interests.

Ethical Approval and Consent to Participate

Ethical approval for this retrospective study was obtained from the Institutional Ethics Committee of Lovely Professional University, Phagwara, India, and permission to access the clinical records was granted by Shrimann Superspeciality Hospital, Jalandhar (Approval No.: LPU/IEC-LPU/2025/1/2; dated 15/02/2025). As the study involved secondary analysis of fully anonymized clinical records, the requirement for written informed consent was waived by the ethics committee. All procedures were performed in accordance with relevant institutional guidelines and the Declaration of Helsinki.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The dataset consists of de-identified clinical data collected from Shrimann Superspeciality Hospital, Jalandhar. Due to patient confidentiality and institutional policies, the data are not publicly available but may be made available upon reasonable request and subject to institutional approval.

Use of Generative AI Tools

During the preparation of this manuscript, the authors used ChatGPT solely for language editing, stylistic refinement, and improvement of structural clarity. All generated content was carefully reviewed, verified, and revised by the authors, who assume full responsibility for the content of the published work.

Author Contributions

Conceptualization: S.M. **Methodology:** S.M., A.S.R. **Software:** S.M. **Validation:** N.M. **Formal Analysis:** S.M. **Data Curation:** A.S.R. **Writing – Original Draft Preparation:** S.M. **Writing – Review & Editing:** N.M., S.K. **Supervision:** N.M.

References

1. World Health Organization, "Dementia: Key facts," 2023, [Internet; cited 2025-09-28]. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia
2. P. Ma, J. Wang, Z. Zhou, C. L. P. Chen, T. A. D. N. I. , and J. Duan, "Development and validation of a deep-broad ensemble model for early detection of alzheimer's disease," *Frontiers in Neuroscience*, vol. Volume 17 - 2023, 2023.
3. L. Yue, W.-g. Chen, S.-c. Liu, S.-b. Chen, and S.-f. Xiao, "An explainable machine learning based prediction model for alzheimer's disease in china longitudinal aging study," *Frontiers in Aging Neuroscience*, vol. Volume 15 - 2023, 2023. [Online]. Available: https://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2023.1267020
4. Y. Wang, S. Liu, A. G. Spiteri, A. L. H. Huynh, C. Chu, C. L. Masters, B. Goudey, Y. Pan, and L. Jin, "Understanding machine learning applications in dementia research and clinical practice: a review for biomedical scientists and clinicians," *Alzheimer's research & therapy*, vol. 16, no. 1, p. 175, 2024.
5. A. G. Vrahatis, K. Skolariki, M. G. Krokidis, K. Lazaros, T. P. Exarchos, and P. Vlamos, "Revolutionizing the early detection of alzheimer's disease through non-invasive biomarkers: the role of artificial intelligence and deep learning," *Sensors*, vol. 23, no. 9, p. 4184, 2023.
6. Z. Jahangir, R. Ranjan, F. Saeed, A. Shiwlani, S. Shiwlani, and M. Umar, "Applications of ml and dl algorithms in the prediction, diagnosis, and prognosis of alzheimer's disease," *American Journal of Biomedical Science & Research*, vol. 22, no. 6, pp. 779–786, 2024.
7. G. Corani, A. Benavoli, J. Demšar, F. Mangili, and M. Zaffalon, "Statistical comparison of classifiers through bayesian hierarchical modelling," *Machine Learning*, vol. 106, no. 11, pp. 1817–1837, 2017.
8. J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.

9. R. D. Riley, K. I. Snell, L. Archer, J. Ensor, T. P. Debray, B. Van Calster, M. Van Smeden, and G. S. Collins, "Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study," *bmj*, vol. 384, 2024.

10. G. S. Collins, K. G. Moons, P. Dhiman, R. D. Riley, A. L. Beam, B. Van Calster, M. Ghassemi, X. Liu, J. B. Reitsma, M. Van Smeden *et al.*, "Tripod+ ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *bmj*, vol. 385, 2024.

11. A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, "Practical guide to shap analysis: Explaining supervised machine learning model predictions in drug development," *Clinical and translational science*, vol. 17, no. 11, p. e70056, 2024.

12. T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

13. L. G. Meteumba, V. P. Ojha, and S. Yarahmadian, "Comprehensive evaluation of machine learning models for predicting the cognitive status of alzheimer's disease subjects and susceptible," *Discover Data*, vol. 3, no. 1, p. 28, 2025.

14. H. A. Helaly, M. Badawy, and A. Y. Haikal, "Deep learning approach for early detection of alzheimer's disease," *Cognitive computation*, vol. 14, no. 5, pp. 1711–1727, 2022.

15. A. Mehmood, A. Abugabah, A. A. AlZubi, and L. Sanzogni, "Early diagnosis of alzheimer's disease based on convolutional neural networks," *Computer Systems Science and Engineering*, vol. 43, no. 1, pp. 305–315, 2022.

16. D. Nguyen, H. Nguyen, H. Ong, H. Le, H. Ha, N. T. Duc, and H. T. Ngo, "Ensemble learning using traditional machine learning and deep neural network for diagnosis of alzheimer's disease," *IBRO Neuroscience Reports*, vol. 13, pp. 255–263, 2022.

17. M. M. S. Fareed, S. Zikria, G. Ahmed, S. Mahmood, M. Aslam, S. F. Jillani, A. Moustafa, M. Asad *et al.*, "Add-net: an effective deep learning model for early detection of alzheimer disease in mri scans," *IEEE Access*, vol. 10, pp. 96 930–96 951, 2022.

18. F. García-Gutierrez, J. Díaz-Álvarez, J. A. Matias-Guiu, V. Pytel, J. Matías-Guiu, M. N. Cabrera-Martín, and J. L. Ayala, "Ga-madrid: Design and validation of a machine learning tool for the diagnosis of alzheimer's disease and frontotemporal dementia using genetic algorithms," *Medical & Biological Engineering & Computing*, vol. 60, no. 9, pp. 2737–2756, 2022.

19. H. Wang, L. Sheng, S. Xu, Y. Jin, X. Jin, S. Qiao, Q. Chen, W. Xing, Z. Zhao, J. Yan *et al.*, "Develop a diagnostic tool for dementia using machine learning and non-imaging features," *Frontiers in aging neuroscience*, vol. 14, p. 945274, 2022.

20. D. Pirrone, E. Weitschek, P. Di Paolo, S. De Salvo, and M. C. De Cola, "Eeg signal processing and supervised machine learning to early diagnose alzheimer's disease," *Applied sciences*, vol. 12, no. 11, p. 5413, 2022.

21. S. Koga, A. Ikeda, and D. W. Dickson, "Deep learning-based model for diagnosing alzheimer's disease and tauopathies," *Neuropathology and Applied Neurobiology*, vol. 48, no. 1, p. e12759, 2022.

22. A. M. Ibrahim, H. R. Abdel-Aziz, H. A. H. Mohamed, D. E. F. Zaghamir, N. M. I. Wahba, G. A. Hassan, M. Shaban, M. El-Nablaway, O. N. Aldughmi, and T. H. Aboelola, "Balancing confidentiality and care coordination: challenges in patient privacy," *BMC nursing*, vol. 23, no. 1, p. 564, 2024.

23.I. Arupzhanov, A. Alimbayev, T. Seyil, T. Aimyshev, T. Maulenkul, A. Oshibayeva, and A. Gaipov, "Methodological note on predicting one-year mortality for chronic diseases using administrative data," *Epidemiology and Health Data Insights*, vol. 1, no. 4, p. ehdi015, 2025.

24.S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

25.R. K. Mohapatra, L. Jolly, and S. P. Dakua, "Advancing explainable ai in healthcare: Necessity, progress, and future directions," *Computational Biology and Chemistry*, p. 108599, 2025.

26.G. Gallitto, R. Englert, B. Kincses, R. Kotikalapudi, J. Li, K. Hoffschlag, U. Bingel, and T. Spisak, "External validation of machine learning models—registered models and adaptive sample splitting," *GigaScience*, vol. 14, p. giaf036, 2025.

27.Y.-Q. Cai, D.-X. Gong, L.-Y. Tang, Y. Cai, H.-J. Li, T.-C. Jing, M. Gong, W. Hu, Z.-W. Zhang, X. Zhang *et al.*, "Pitfalls in developing machine learning models for predicting cardiovascular diseases: challenge and solutions," *Journal of Medical Internet Research*, vol. 26, p. e47645, 2024.