*Proceedings of the Pakistan Academy of Sciences: A Physical and Computational Sciences* 62(4): 323-336 (2025)
Copyright © Pakistan Academy of Sciences
ISSN (Print): 2518-4245; ISSN (Online): 2518-4253
http://doi.org/10.53560/PPASA(62-4)696

Pakistan Academy of Sciences

Research Article

# Hybrid Supervised Machine Learning Models for Enhanced Alzheimer's Disease Classification

**Muazzam Ali[1*], M.U. Hashmi[1], Zakeesh Ahmad[2], Noor Ul Ain Kazmi[2], Asifa Ittfaq[2], and Amna Ashraf[2]**

[1*]Department of Computer Sciences, Superior University, Lahore, Pakistan
[2]Department of Basic Sciences, Superior University, Lahore, Pakistan

**Abstract:** This research aims to facilitate the early and precise identification of Alzheimer's disease (AD), which remains one of the most prevalent neurodegenerative diseases impacting people's health and quality of life around the world. Employing machine learning algorithms, this study aims to develop reliable and effective models that support clinical workflows and streamline processes, thereby reducing the burden on patients and their families and ultimately enhancing patient-centric diagnostic frameworks. An approach to data cleaning, involving data imputation, encoding categorical variables, normalization of certain features, and stratified training and testing data splitting with hyperparameter tuning, was employed. This approach utilized both grid search and stratified k-fold cross-validation. Traditional models, ensemble techniques, and hybrid models were tested, including Lasso + LightGBM, XGBoost + SVM, and blended models such as LightGBM, CatBoost, Logistic Regression, and XGBoost. Lasso + LightGBM outperformed others in hybrid models. Lasso + LightGBM achieved an accuracy of 0.961240, precision of 0.943231, recall of 0.947368, and F1score of 0.945295, Cohen's Kappa of 0.915284, Hamming Loss of 0.038760, and Jaccard Index with the value of 0.896266. This research contributes to UNSDG 3, "Good Health and Well-being", by enhancing data-driven health education and resources, and an efficient diagnostic and management system for Alzheimer's. It also promotes healthy aging globally among the population.

**Keywords:** Predictive Modeling, Biomedical Data Analysis, Feature Engineering, Gradient Boosting, Clinical Decision Support, Cross-Validation, Diagnostic Accuracy.

## 1. INTRODUCTION

Alzheimer's is a behavior and progressive dementia disorder that impacts behavior, and thinking to a major extent and memory. It is the most common form of dementia, which induces tremendous loss of cognitive ability as people grow older [1]. Diagnosing Alzheimer disease is challenging as it can resemble the aging process or other brain-related diseases. In modern times, diagnosis is made through cognitive tests, brain scans, as well as clinical examinations, which are subjective and time-consuming [2, 3]. It does not have a single conclusive test, which is why detecting it early is a challenge, as it is crucial to the treatment and management of the condition. Following advances in machine learning (ML), a potent tool has emerged for enhancing the diagnosis of Alzheimer's disease by analyzing large and complex medical data. Patterns in the patient data have been drawn using traditional statistical methods and simple ML algorithms like the Naive Bayes and K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). These methods produce fast results; however, as with high-dimensional data, such as brain scans and genetic data, the methods are not particularly effective, which restricts their accuracy [4]. Ensembles and deep learning are sophisticated machine learning methods that help to mitigate these challenges. Cloud random Forests and gradient boosting are ensemble models that involve using a combination of models to improve the accuracy of predictions [5-7]. Deep learning-based models, such as Convolutional Neural Networks

Received: July 2025; Revised: November 2025; Accepted: December 2025

* Corresponding Author: Muazzam Ali <muazzamali@superior.edu.pk>

(CNNs), are indeed powerful tools that enable the processing of medical images and the detection of subtle changes in medical imaging (e.g., MRI, PET) associated with Alzheimer's disease. It is possible to improve patient outcomes by enhancing diagnosis accuracy and reducing the diagnosis period, thereby decreasing the risks of human error and leading to a better situation for clients. By providing better, more accurate, and timely diagnostics, researchers will be able to improve both treatment strategies and disease prevention [8, 9].

Current techniques of Alzheimer's disease (AD) diagnosis predominantly focus on genetic factors that involve machine learning and deep learning models, particularly by analyzing gene expression data for early detection of the disease. Studies have shown that deep learning (DL) models, including DGS-TabNet, outperform traditional ML algorithms by selecting more precise and efficient meaningful genes, obtaining superior classification performance (up to 93.8% accuracy and 98.53% Area under Curve (AUC) in binary classification tasks). Moreover, some key genes may also have biological significance by revealing their roles in other diseases, which could partly confirm that the use of network-based analyses in conjunction with traditional methods is valuable for identifying genetic markers related to AD [10]. Alzheimer's disease prediction has been significantly enhanced by recent machine learning algorithms, particularly those utilizing ensemble models (e.g., LightGBM and Random Forest), which can achieve accuracies exceeding 96.35% on several databases [11]. The use of Shapley Additive Explanation (SHAP) and Local Interpretable Model-agnostic Explanation (LIME) enhances artificial intelligence (AI) explainability, and as a result, the model's transparency leads to higher clinician trust in it. Compared to existing methods that are restricted by the number of datasets, data type, or interpretability, this method has improved efficiency and usability in AD diagnosis [12]. Mahamud *et al*. [13] developed a framework that uses data on handwriting to detect Alzheimer's disease, which involves a two-phase forward-backward selection of features via XGBoost. This strategy limits the workflow to a minimal set of tasks to increase interpretability to achieve 91.37% accuracy. The robust performance by using the leave-one-out cross-validation indicates that the sample size was adequate and transforms towards more friendly AD diagnosis. The present study also provides autography as a more reliable and straightforward strategy for early detection of AD.

The proposed research problem in the present study is the Computer-Aided Diagnosis (CAD) of Alzheimer's disease, which is addressed by designing and testing hybrid supervised machine learning models that combine adaptive feature selection, blended probability fusion, and gradient boosting. Responses to existing research have proven encouraging with the use of individual classifiers and the simple ensemble technique; however, they often fail to address high-dimensional, imbalanced, and heterogeneous clinical data, which ultimately results in poor generalizability and reduced clinical interpretability. To address these weaknesses, this work generalizes gradient boosting in a meta-modeling system, which has enhanced the robustness, discrimination, and interpretability of both linear and nonlinear learners.

The dataset used in the present study is the result of less controlled environments, specifically community-based and non-specialist clinical environments, where the data may be noisier, less standardized, and even completely missing, compared to strictly controlled research studies. This feature drove the adoption of hybrid designs that can tolerate uncertainty and variability while preserving the performance of diagnosis. In this connection, the objectives of this study will be the following:

- To build and test a set of hybrid machine learning models to classify Alzheimer's disease, which incorporate feature selection (i.e., Lasso) with effective gradient-boosting algorithms (i.e., LightGBM, XGBoost, CatBoost).
- To evaluate the capabilities of such hybridization in terms of predictive reliability and robustness, in comparison with standalone methods and conventional ensemble methods reported in recent literature.
- To ensure that the final models can be interpreted clinically, where interpretability is measured by the sparsity of the chosen features and the transparency of the linear elements in the hybrid structures.

The present study focuses on integrating and benchmark existing strategies to address the issue in the Alzheimer's CAD system. These issues include data heterogeneity, small sample size and transparency of the model. Rather than proposing the new model, the approach in the present study aims to increase the effectiveness of current models, by developing the ML models that are clinically viable and applicable in practice.

## 2. METHODOLOGY

### 2.1. Dataset and Preprocessing

The Alzheimer's disease dataset, which was submitted to Kaggle by Rabie El Kharoua in 2024 and is released under the Attribution 4.0 International (CC BY 4.0) license (DOI: 10.34740/KAGGLE/DSV/8668279), is utilized in this research. 35 variables, including demographic, lifestyle, medical history, cognitive evaluation, symptoms, and diagnostic information pertaining to Alzheimer's disease, are included in the dataset, which includes 2,149 patient records (IDs 4751-6900). Because it is a binary variable that indicates whether Alzheimer's disease is present (1) or not (0), the diagnosis column is the target variable.

### *2.1.1. Handling missing values*

Missing values in the dataset can compromise the reliability of model predictions. Therefore, all missing data are imputed using the mode (i.e., the most frequent value) for each column [14]. This approach is mathematically expressed as:

$$\hat{q}_i = mod(q_{i,1}, q_{i,2}, q_{i,3}, \dots \dots \dots \dots q_{i,n}) \quad (1)$$

Where $\hat{q}_i$ denotes the imputed value for feature i, while n represents samples. This method ensures the categorical and numerical integrity of the dataset, preserving both the sample size and variance structure.

### *2.1.2. Categorical encoding*

To transform categorical variables into a numerical format, Label Encoding is applied to all features except the target column [15]. Each category is mapped to a unique integer, enabling the models to process categorical features mathematically:

$$\text{Encoded}(x) = i, \text{where } x \in \text{Categories}, i \in N \quad (2)$$

### *2.1.3. Normalization*

For all continuous features, normalization using the Standard Scalar is performed, transforming the data to have a zero mean and unit variance [16].

$$z = \frac{x-\mu}{\sigma} \quad (3)$$

where σ is the standard deviation, μ is the mean, and x is the initial value for each feature. To guarantee that feature-scaling-sensitive models (like SVM and KNN) operate at their best, this step is essential.

### *2.1.4. Feature importance*

The features of the Alzheimer's disease dataset have been ranked based on the scores of feature importance from the model using Random Forests, as illustrated in Figure 1. Random Forest has been used because the dataset is not very large, and it is capable of handling a large number of features without any problem. Functional Assessment and ADL (Activities of Daily Living) were the

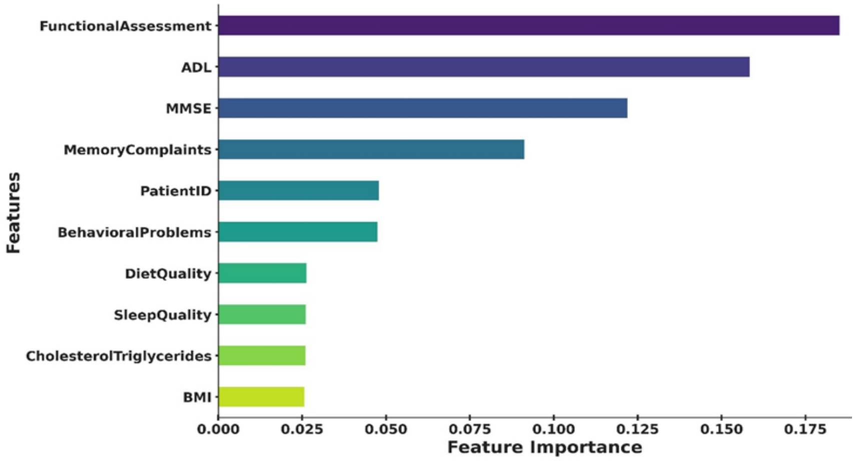

**Fig. 1.** Top 10 feature importance for Alzheimer's classification.

most significant factors. Therefore, they are the most important in predicting whether a case is Alzheimer's disease or non-Alzheimer's disease.

The other characteristics, such as the MMSE (Mini-Mental State Examination) and Memory Complaints, also play a significant role, showing that they are important in the clinical assessment of cognitive abilities. Conversely, the importance of features such as Cholesterol/Triglycerides, Sleep Quality, and Diet Quality is lower, which is a sign of weakness in these variable predictors in the dataset. This distribution is logical, given that functional and cognitive assessments are primary constituents for diagnosing Alzheimer's disease, thereby confirming the dataset's primary clinical relevance.

## 2.2. Data Splitting

A stratified train-test split is utilized to maintain class distribution in both sets. 70% of the data is allocated for training (Xtrain,ytrainX_{train}, y_{train}Xtrain,ytrain), and 30% for testing (Xtest,ytestX_{test}, y_{test}Xtest,ytest), ensuring that performance metrics generalize to unseen data.

$$(X_{train}, Y_{train}), (X_{test}, Y_{test}) = \mathrm{StratifiedSplit}(X, Y, test_{score} = 3.0) \quad (4)$$

## 2.3. Model Training and Hyperparameter Tuning

A variety of supervised learning models are compared, with a particular focus on hybrid models developed by combining model outputs or feature selection pipelines. We performed hyperparameter optimization using GridSearchCV with stratified k-fold cross-validation (k = 5) to optimize precision and recall. We aimed to optimize the F1-score as the basic criterion for the model selection. In this process, stratified fold cross-validation was used to preserve the properties of class, decreasing the risk of overfitting. Moreover, this strategy ensured that hyperparameter estimation remains robust.

## 2.4. Used Models

We trained models using grid search with traditional classifiers, including Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, and boosting and bagging techniques (XGBoost, LightGBM, CatBoost, AdaBoost, and Bagging Classifier). These may be used as standalone benchmarks or in conjunction with hybrid approaches. The model parameters are listed in Table 1.

### *2.4.1. K-nearest neighbors (KNN)*

KNN is a non-parametric, instance-based algorithm where classification is based on the majority vote among the k closest training samples in the feature space [17]. The value of k is selected via grid search. The distance metric, typically Euclidean, is calculated as:

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{p} (x_{il} - x_{jl})^2} \quad (5)$$

The size of the data affects this approach; hence, the previously mentioned normalization step is required. The curse of dimensionality can cause KNN's performance to deteriorate in high-dimensional environments, yet it is still a useful baseline for tabular datasets with modest complexity [18].

### *2.4.2. AdaBoost*

Adaptive Boosting, also known as AdaBoost, is a technique that builds a powerful classifier by repeatedly training weak learners, typically decision stumps. However, each new learner is modeled after its predecessors, focusing on their mistakes [19]. The last model is the weighted sum of such learners:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \quad (6)$$

Where $\alpha_t$ is the weight assigned to weak classifier $h_t(x)$. AdaBoost is especially robust to overfitting in many practical cases, but can be sensitive to noisy data and outliers.

### *2.4.3. Bagging (bootstrap aggregating)*

Bagging trains multiple base estimators on different bootstrap samples of the dataset and averages their predictions to reduce variance. For binary classification:

$$\hat{y} = majority\_vote(h_1(x), h_1(x), \ldots, h_n(x)) \quad (7)$$

This strategy makes the models more stable especially, when using high variance base

**Table 1.** Hyperparameters tuned and their grid search values for each machine learning model.

| Model | Hyperparameter Name | Hyperparameter Values |
|---|---|---|
| RandomForestClassifier | n_estimators, max_depth, min_samples_split, min_samples_leaf, bootstrap | 100, 10, 2, 1, True |
| SVM (Support Vector Machine) | C, kernel, gamma, degree, coef0, tol | 1, rbf, scale, 3, 0.0, 1e-3 |
| KNN (K-Nearest Neighbors) | n_neighbors, weights, algorithm, leaf_size, p | 5, uniform, auto, 30, 2 |
| LogisticRegression | C, penalty, solver, max_iter, tol | 1, l2, lbfgs, 100, 1e-3 |
| XGBoost | n_estimators, learning_rate, max_depth, subsample, colsample_bytree, gamma | 100, 0.1, 6, 0.8, 0.8, 0.1 |
| LightGBM | n_estimators, learning_rate, max_depth, num_leaves, min_child_samples, subsample | 100, 0.1, 6, 31, 20, 0.8 |
| CatBoost | iterations, learning_rate, depth, l2_leaf_reg, subsample, colsample_bylevel | 100, 0.1, 6, 3, 0.8, 0.8 |
| AdaBoost | n_estimators, learning_rate, algorithm | 100, 1.0, SAMME.R |
| Bagging | n_estimators, max_samples, max_features, bootstrap, n_jobs | 100, 1.0, 1.0, True, -1 |
| StackingClassifier | estimators, final_estimator, cv | RandomForestClassifier, XGBClassifier, LogisticRegression, 5 |
| RF + Logistic Regression (Stacked) | rf__n_estimators, rf__max_depth, rf__min_samples_split, rf__min_samples_leaf, lr__C, lr__penalty, lr__solver | 100, 10, 2, 1, 1, l2, lbfgs |
| XGBoost + SVM (Stacked) | xgb__n_estimators, xgb__learning_rate, xgb__max_depth, svm__C, svm__kernel, svm__gamma | 100, 0.1, 6, 1, rbf, scale |
| Lasso + LightGBM (Hybrid) | lasso__alpha, lgbm__n_estimators, lgbm__learning_rate, lgbm__max_depth, lgbm__num_leaves, lgbm__min_child_samples | 0.1, 100, 0.1, 6, 31, 20 |
| RF-FeatureSelection + LR (Hybrid) | rf__n_estimators, rf__max_depth, rf__min_samples_split, rf__min_samples_leaf, lr__C, lr__penalty, lr__solver | 100, 10, 2, 1, 1, l2, lbfgs |
| Blended Probabilities (LGBM + CatBoost + XGB) + LR | lgbm__n_estimators, lgbm__learning_rate, catboost__iterations, catboost__learning_rate, xgb__n_estimators, lr__C | 100, 0.1, 100, 0.1, 100, 1 |

This rigorous methodology underpins both the fairness and scientific validity of model comparison, ensuring that reported results are robust, replicable, and meaningful for biomedical decision-making.

learners like decision trees. Hyperparameters (e.g. estimators) are optimized with the help of cross-validation [20].

### *2.4.4. Logistic regression*

Standard Logistic Regression is used as a core linear baseline [21]. It estimates the probability of the binary outcome using the logistic function:

$$P(y = 1 \mid x) = \frac{1}{1 + exp(-(\beta_0 + \beta^T x))} \quad (8)$$

When the coefficients of $\beta$ are estimated using the maximum likelihood method. C is a parameter that is regularized to control the model's complexity. Despite being linear, Logistic Regression is likely to compete with biomedical data and provide understandable coefficients.

It is not new to use some of the models employed in the present study; however, when applied to a comparatively strict and data-driven technique for Alzheimer's disease, which has high dimensionality and noise, they are instructive in science. Not merely accumulating, but this choice supposes the potential of an orderly examination of model action and hybrid synergy, by which empirically information on what architectures will be evident in most clinically diverse situations. This is the gap, which is negatively addressed in the literature.

## 2.5. Hybrid Model Architectures

The study constructs and evaluates five advanced hybrid models, each leveraging the strengths of its constituent algorithms to address the nonlinearity, feature interaction, and potential collinearity within the dataset.

### *2.5.1. Hybrid 1: Random forest probabilities as features for logistic regression (RF + LR)*

First, a Random Forest classifier is trained on the original feature set, outputting class probabilities for each sample:

$$P_{RF}(y = 1 \mid x) = \frac{1}{n_{trees}} \sum_{t=1}^{n_{trees}} h_t(x) \tag{9}$$

Where $h_t(x)$ is the prediction probability from tree t. The predicted probability $P_{RF}$ is then appended as a new feature to both the training and test datasets:

$$X' = [X, P_{RF}] \tag{10}$$

The hybrid RF+LR model follows a two-stage stacking formulation. Consider $f_{RF}(x)$ is the random forest probability estimator then:

$$f_{RF} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{11}$$

We produce out of fold (OOF) predictions by using:

$$\hat{p}_{RF,i} = f_{RF}^{(-k)}(x_i) \tag{12}$$

The meta feature matrix becomes:

$$X^{RF} = [X, \hat{p}_{RF}] \tag{13}$$

Now the logistic regression function for the decision is given by:

$$f_{LR}(X^{RF}) = \sigma(\beta_0 + \beta^T X + \gamma \hat{p}_{RF}) \tag{14}$$

$\gamma$ represent the weight assigned to RF-derived probability, so the final hybrid prediction is computed using:

$$\hat{y} = 1\{f_{RF}(X^{RF}) > 0.5\} \tag{15}$$

A Logistic Regression model is subsequently trained on X′, learning a linear boundary in the enriched feature space. This hybridization combines the nonlinear feature extraction capability of Random Forests with the interpretability and regularization strength of Logistic Regression. The hybrid model can potentially address nonlinearity and feature interactions missed by Logistic Regression alone. However, there is a risk of overfitting if the new probability feature is highly correlated with the target, particularly in small or unbalanced datasets. In this study, cross-validation and the use of the test set mitigate such risks [22].

### *2.5.2. Hybrid 2: XGBoost probabilities as features for SVM (XGBoost + SVM)*

An XGBoost model, known for its gradient-boosted tree structure and robustness to feature collinearity, is first trained. The predicted probabilities for each sample, $P_{XGB}$, are calculated:

$$P_{XGB}(y = 1 \mid x) = \sigma(f_{XGB}(X)) \tag{16}$$

Where σ denotes the sigmoid function. These probabilities are appended as an additional feature to the input matrix, after which a Support Vector Machine (SVM) classifier is trained and OFF probabilities $\hat{p}_{XGB}$ are concatenated with the input features:

$$X'' = [X, P_{XGB}] \tag{17}$$

The SVM with a radial bases function (RBF) kernel learns separating hyperplane in the augmented space:

$$f_{SVM}(X'') = sign(w^T X) + \gamma \hat{p}_{XGB} + b \tag{18}$$

The term $\gamma \hat{p}_{XGB}$ quantifies the contribution of initial stage boosted the probabilities to SVM margin. This hybrid combines XGBoost's nonlinear learning capacity with the margin-maximizing properties of SVMs. This approach can significantly enhance performance if XGBoost probabilities encapsulate a high-level structure that is not easily captured by SVM alone. However, SVMs are sensitive to irrelevant features, so the benefit depends on the informativeness of the probability feature [23].

### *2.5.3. Hybrid 3: Lasso feature selection followed by LightGBM (Lasso + LightGBM)*

A Logistic Regression model with L1 regularization (Lasso) is employed to perform feature s A Logistic Regression model with $L_1$ regularization (Lasso) is employed to perform feature selection:

$$min_{\beta} = (-logL(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|) \quad (19)$$

Where $L(\beta)$ is the likelihood, $\beta_j$ are the coefficients, and $\lambda$ is the regularization parameter. Only features with nonzero coefficients are retained:

$$S = \{j : \beta \neq 0\} \quad (20)$$

The reduced feature matrix is:

$$X^{Lasso} = X[:, S] \quad (21)$$

LightGBM is trained on the reduced space:

$$\hat{y} = f_{LGBM}(X^{Lasso}) \quad (22)$$

This hybrid is a sequential architecture an optimizing based selector followed by the gradient boosting. LightGBM, a fast and efficient gradient boosting implementation, is trained on the selected features. This hybrid is especially effective in high-dimensional data, as it removes redundant and noisy variables before applying a strong tree-based learner. The risk is that overly aggressive feature selection can discard weak but informative features, potentially lowering overall model capacity [24].

### *2.5.4. Hybrid 4: Top N random forest feature importance with logistic regression (RF-Feature Selection + LR)*

Random Forests naturally provide feature importance measures based on mean decrease in impurity (MDI) or mean decrease in accuracy (MDA). Random forest computed the importance values by:

$$I_j = \sum_{t=1}^{T} \sum_{s \in S_{t,j}} \Delta i(s) \quad (23)$$

The top $N$ features with the highest importance scores are selected:

$$S_N = argsort(Importance_{RF})[:N] \quad (24)$$

Logistic regression is trained on:

$$X^{RF} = X[:, S] \quad (25)$$

The model is then given by:

$$f_{LR}(X^{RF}) = \sigma(\beta_0 + \beta^T X^{RF}) \quad (26)$$

This hybrid is featuring selection driven linear model contrasting with fully nonlinear boosters. Logistic Regression is then trained on this reduced feature set. Selecting the most predictive variables reduces dimensionality and may improve generalization, especially for linearly separable relationships. However, feature importance scores can be unstable in the presence of multicollinearity or redundant predictors, and choosing N is somewhat heuristic [25].

### *2.5.5. Hybrid 5: Blended probabilities of multiple boosting models with logistic regression (Blended Probabilities + LR)*

LightGBM, CatBoost, and XGBoost models are independently trained on the original dataset. For each sample, the predicted probabilities from each model are extracted:

$$P_{LGBM}(y = 1 \mid X) \quad (27)$$

$$P_{CAT}(y = 1 \mid X) \quad (28)$$

$$P_{XGB}(y = 1 \mid X) \quad (29)$$

These probabilities are concatenated with the original features to create a new, augmented feature space:

$$X''' = [X, P_{LGBM}, P_{CAT}, P_{XGB}] \quad (30)$$

Let the blended meta feature vector be:

$$z_i = \begin{bmatrix} P_{LGBM,i} \\ P_{CAT,i} \\ P_{XGB,i} \end{bmatrix} \quad (31)$$

The final model is:

$$f_{LR}(X'') = \sigma(\beta^T X + \alpha_1 P_{LGBM} + \alpha_2 P_{CAT} + \alpha_3 P_{XGB} + b) \quad (32)$$

This is the probabilistic blending architecture that combines diverse gradient boosting models.

A Logistic Regression model is trained on X‴, learning how to combine the output of diverse boosting models optimally. This method synthesizes predictions from heterogeneous boosting frameworks, enabling the final model to exploit differences in model behavior [26]. While potentially powerful, this approach increases the risk of overfitting if the boosting models themselves are highly correlated or overfit the training data.

The benefits of these hybrid models extend beyond the advantages of conventional classifiers (such as Random Forest and Logistic Regression) to more complex algorithms, including feature selection with Lasso, boosting on XGBoost, LightGBM, and CatBoost, as well as ensemble learning methods like Stacking and Blended Probabilities. The hybrid models that use the probabilities generated by one model as input for the other model are helpful for the consideration of complexities like intricate feature interactions and nonlinearity that providing a novel approach to increase the model performance. A stronger decision is achieved using combined models, such as RF + LR, XGBoost + SVM, and Lasso + LightGBM, which present a novel perspective for processing high-dimensional imbalanced data.

### 2.6. Evaluation Metrics

The performance and robustness of these classification models are evaluated using specific metrics. These provide complementary information, accurately reflecting the overall correctness of the model, while precision measures how many of the predicted positives are truly positive. Recall shows how many actual positives are identified correctly and the F1-score balances the tradeoff between false positive and false negative. Cohen's Kappa, Hamming loss, and Jaccard Index capture the nuances of agreement and multi-label performance. The use of these measures enables a more advanced and less biased assessment of predictive models in various situations under different data distributions [27].

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{33}$$

$$precision = \frac{TP}{TP + FP} \tag{34}$$

$$recall = \frac{TP}{TP + FN} \tag{35}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{36}$$

$$Cohen's\ Kappa = \frac{P_0 - P_e}{1 - P_e} \tag{37}$$

$$Hamming\ Loss = \frac{1}{N}\sum_{i=1}^{N} 1(y_i \neq \hat{y}_i) \tag{38}$$

$$Jaccard\ Index = \frac{|A \cap B|}{|A \cup B|} \tag{39}$$

This rigorous methodology underpins both the fairness and scientific validity of model comparison, ensuring that reported results are robust, replicable, and meaningful for biomedical decision-making.

## 3. RESULTS AND DISCUSSION

The cross-evaluation of model benchmarks reveals reasonable differences in various measures, indicating the impact of different machine learning and hybrid methods for classifying Alzheimer's disease. Table 2 presents the evaluation metrics values for all models. The best accuracy is reported for , CatBoost, and Lasso + LightGBM, both scoring 0.961240, closely followed by XGBoost 0.961041, LightGBM and stacking at 0.958140 and Blended Probabilities (LGBM + CatBoost + XGB) + LR at 0.956589. This identifies the better performance of gradient boosting-based and ensemble hybrid methods for classifying the disease status. On the other hand, the KNN (0.737984) and RF-FeatureSelection + LR (0.846512) models exhibit relatively lower accuracy, which stems from high dimensionality and the sensitivity to feature selection, respectively. The accuracy achieved in this research is slightly higher than previous values of 0.9635 reported by Mahamud *et al.* [13] and 0.9380 recorded by Jin *et al.* [10].

Table 2 shows that the highest precision is recorded for CatBoost (0.951111) and XGBoost + SVM (0.950893), which are higher than the previous values of 0.95 stated by Mahamud *et al.* [13] and 0.9396 (with proposed model), reported by Jin *et al.* [10]. Both are effective in minimizing false positive rates and thereby curtailing diagnosis overestimation, which is crucial for less invasive procedures in clinical practice. Traditional classifiers, such as SVM (0.774336) and KNN (0.680982), perform markedly worse and are often unable to manage the class imbalance and complexity of features, despite normalization.

**Table 2.** Performance metrics (Accuracy, Precision, Recall, F1-Score) for various machine learning models evaluated in Alzheimer's disease classification.

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| RandomForest | 0.941085 | 0.943925 | 0.885965 | 0.914027 |
| SVM | 0.838760 | 0.774336 | 0.767544 | 0.770925 |
| KNN | 0.737984 | 0.680982 | 0.486842 | 0.567775 |
| LogisticRegression | 0.838760 | 0.787037 | 0.745614 | 0.765766 |
| XGBoost | 0.961041 | 0.941831 | 0.943468 | 0.941155 |
| LightGBM | 0.958140 | 0.938865 | 0.942982 | 0.940919 |
| CatBoost | 0.961240 | 0.951111 | 0.938596 | 0.944812 |
| AdaBoost | 0.927132 | 0.891775 | 0.903509 | 0.897603 |
| Bagging | 0.947287 | 0.925439 | 0.925439 | 0.925439 |
| Stacking | 0.958140 | 0.942731 | 0.938596 | 0.940659 |
| RF + LR | 0.945736 | 0.940639 | 0.903509 | 0.921700 |
| XGBoost + SVM | 0.959690 | 0.950893 | 0.934211 | 0.942478 |
| Lasso + LightGBM | 0.961240 | 0.943231 | 0.947368 | 0.945295 |
| RF-FeatureSelection + LR | 0.846512 | 0.781659 | 0.785088 | 0.783370 |
| Blended Probabilities (LGBM + CatBoost + XGB) + LR | 0.956589 | 0.942478 | 0.934211 | 0.938326 |

The XGBoost and Lasso + LightGBM achieved the highest values of recall, that are 0.943468 and 0.947368, respectively, that is higher than the value of 0.9380, reported by Jin *et al*. [10]. This aspect is crucial in clinical practice, where this kind of performance is needed to minimize the number of missed cases. Models such as the KNN model (score = 0.486842) have vast potential for further improvement, indicating that a simple model is underfitted in the presence of complex data.

As shown in Table 2, XGBoost (0.941155), CatBoost (0.944812), and Lasso + LightGBM (0.945295) achieved the highest F1-score, indicating that they can balance the precision-recall tradeoff better than other models, which is crucially important for medical diagnosis. The error spread is small; therefore, we can expect good accuracy from these algorithms.

Table 3 presents the Cohen's Kappa values of hybrid and ensemble approaches, including XGBoost (0.915284), CatBoost (0.914946), and Lasso + LightGBM (0.915284), which demonstrate considerable reliability in model classification consistency and performance, as well as reasonable performance. With Kappa point classification, the SVM (0.646527) and KNN (0.387154) are considered too soft, indicating that both have insufficient reliability to validate incomplete agreement. The hamming loss value is decreased with perfect classification and is particularly low when models XGBoost (0.038760), Catboost (0.038760) and Lass + LightGBM (0.038760) outperform the other models. As expected, KNN, due to its loss, suffers significant losses, which remain at 0.262016, primarily due to poor recall and precision, resulting in numerous mismatches. The three algorithms, XGBoost, CatBoost, and Lasso + LightGBM, scored the best with scores of 0.896266, 0.895397, and 0.896266, respectively, indicating that they have better predictive ability than other models and align more closely with the predicted true label. Many traditional and hybrid strategies like KNN (0.39642) and RF-feature Selection + LR (0.64388) performed below the chance level as expected due to their lower overall classification performance.

These results support the reasoning behind the methodology's focus on ensembles of hybrid models, as the integration of feature selection with probabilistic augmentation and gradient boosting is expected to improve performance significantly. The dataset underwent extensive preprocessing, including the meticulous imputation of missing values, label encoding, normalization, and stratified train-test splitting, which preserved class

**Table 3.** Cohen's Kappa, Hamming Loss, and Jaccard Index scores for different machine learning models in Alzheimer's disease classification.

| Model | Cohen Kappa | Hamming Loss | Jaccard Index |
|---|---|---|---|
| RandomForest | 0.869284 | 0.058915 | 0.841667 |
| SVM | 0.646527 | 0.161240 | 0.627240 |
| KNN | 0.387154 | 0.262016 | 0.396429 |
| LogisticRegression | 0.642971 | 0.161240 | 0.620438 |
| XGBoost | 0.915284 | 0.038760 | 0.896266 |
| LightGBM | 0.908506 | 0.041860 | 0.888430 |
| CatBoost | 0.914946 | 0.038760 | 0.895397 |
| AdaBoost | 0.841049 | 0.072868 | 0.814229 |
| Bagging | 0.884671 | 0.052713 | 0.861224 |
| Stacking | 0.908324 | 0.041860 | 0.887967 |
| RF + LR | 0.880208 | 0.054264 | 0.854772 |
| XGBoost + SVM | 0.911455 | 0.040310 | 0.891213 |
| Lasso + LightGBM | 0.915284 | 0.038760 | 0.896266 |
| RF-FeatureSelection + LR | 0.664523 | 0.153488 | 0.643885 |
| Blended Probabilities (LGBM + CatBoost + XGB) + LR | 0.904834 | 0.043411 | 0.883817 |

proportions to ensure the data's integrity while enhancing model generalizability. Grid search with stratified cross-validation for class-preserved folds enabled extensive multi-criteria hyperparameter optimization, minimizing the risk of overfitting and further augmenting model performance through fine-tuned hyperparameter adjustment.

The complicated nonlinear correlations observed in clinical and demographic data cannot be fully represented by simpler models such as KNN and Logistic Regression, in addition to the more traditional boundary-defining approximations and closest neighbor assumptions. The successful use of feature engineering and hyperparameter tuning has led to the development of clinical decision support tools for testing, highlighting the potential of complicated ensemble models for early Alzheimer's disease identification.

Figure 2 illustrates the pairwise distributions and interrelations between the significant predictors (Functional Assessment, ADL, MMSE, Memory Complaints, Behavioral Problems, and Sleep Quality) by diagnosis class. It is also easy to note clear differences between the Alzheimer and non-Alzheimer groups of the Functional Assessment, ADL, and MMSE, which indicates their great discriminative power. Contrastingly, Memory Complaints and Behavioral Problems have a higher overlap, meaning a lower predictive ability independently. Such visual trends are reflected in the rankings of feature importance gained with the help of Random Forest and Lasso selection, with functional and cognitive measures prevailing. Feature selection methods like mRMR and mutual information have also explained their efficiency in enhancing the prediction of Alzheimer's disease with an accuracy of 0.9908 [28].

More importantly, the figure also presents qualitative data on why the hybrid and ensemble models (e.g., Lasso + LightGBM) performed well: these models can learn nonlinear and partially collinear relationships between features, especially between cognitive and behavioral variables. Such curved or overlapping boundaries are not easily modeled using standard linear classifiers (e.g., Logistic Regression), which is why such classifiers achieve relatively low recall and F1 scores. That is why a pair-plot is not only justifying feature selection, but also the models' success, as it sets up the data structure visually and demonstrates where simple models may fail.

### 3.1. Model Behavior and Error Analysis

Lasso + LightGBM. L1 selection yielded a sparse

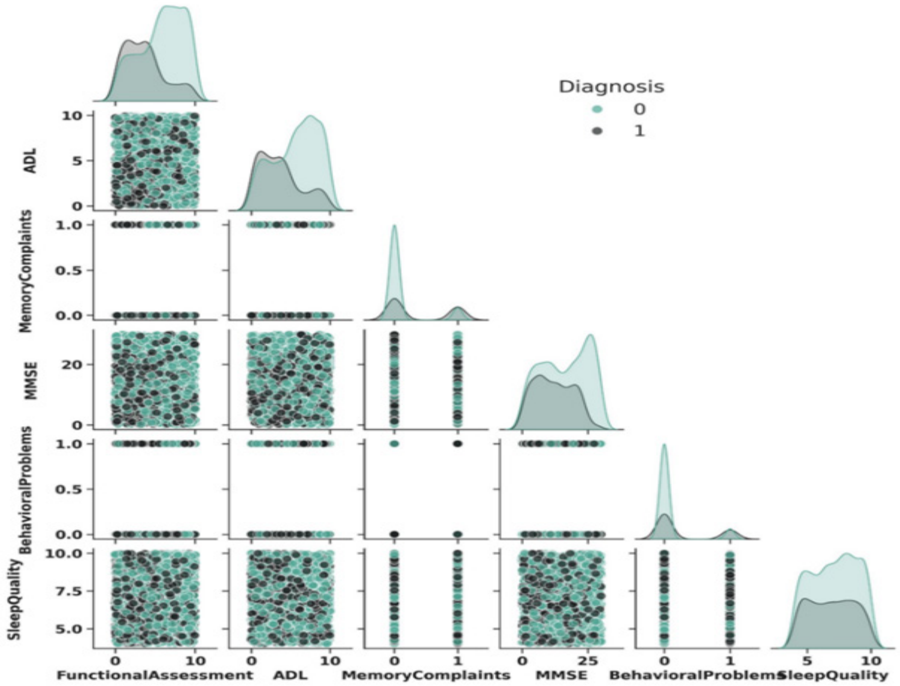

**Fig. 2.** Pairwise feature relationships by Alzheimer's diagnosis.

and lower-correlation subsample that reduces noise and redundancy; LightGBM then learned nonlinear interactions in this low-dimensional space, which are consistent with the more evident separations in the cases of Functional Assessment, ADL, and MMSE in Figure 2. Blended probabilities + LR. The base boosters had a high prediction correlation due to theoretical gains, which constrained the meta-learner's ability to be diverse. In a small sample size, the inclusion of correlated probability enhanced variance and decreased net benefit; moreover, variation in probability calibration was likely a restraining factor for the LR combiner. The combination of RF with Adaboost achieved 0.9255 accuracy which explained the benefits of ensemble learning in boosting model performance. The combination of DT, Adaboost and LR achieved highest accuracy of 0.9546 which shows the effectiveness of blending different models [29].

The study relies on a single dataset from Kaggle that may limit the generalizability of the model to clinical datasets. The models in the present study were evaluated only on provided dataset and external validation on an independent dataset was not performed. It is difficult to confirm the robustness and real-world applicability of the proposed models. Hybrid models such as Lasso + LightGBM and blended probabilities show strong performance; these may remain complex and less interpretable. This can limit their practical use in clinical settings where model transparency and interpretability are very important for clinical trust and decision-making.

RF-FeatureSelection + LR. RF importances based on impurity can be unstable under collinearity and biased against specific types of features; in a top-N heuristic, weak yet informative variables can be discarded. A linear LR fitted on this subset underfits the nonlinear structure, shown in Figure 2, which explains the gap between the accuracy and recall. Practical note: Future variants will (i) apply permutation/Boruta or stability selection to features, (ii) impose out-of-fold predictions and temperature/Platt calibration in blending, and (iii) take into account Elastic-Net LR or monotone-constrained boosting to make the thus far observed structure more like reality.

To evaluate the robustness, consistency and adaptability of the models, we used many established mechanisms. Robustness and generalization were assessed by using the stratified 5-folds cross validation, where models were trained and validate on multiple class preserving split ad by using was the out-of-fold (OOF) predictions to avoid the information leakage in hybrid stacking. Consistency was verified by using a various set of metrics like accuracy, precision, recall, F1score, kappa, hamming loss, Jaccard index that showed stable rankings within the Table 2 and Table 3. Adaptability was evaluated testing the models on heterogeneous mix of demographic. Cognitive, behavioral and clinical features. Lastly, all results were confirmed on a held 30% unseen test set to ensure the valid generalization.

### 3.2. Comparative Discussion

Direct and cross-paper comparisons of point estimations (e.g., accuracy or F1) are necessarily constrained since results are highly dependent on the particular dataset (size, difficulty, feature set, and class balance) and preprocessing options, as well as the evaluation protocol. We therefore do not claim that we are better than previous studies solely because our point estimates (e.g., accuracy 0.961) are numerically larger than those obtained with other datasets and setups (e.g., 0.938). Rather, we place our findings on a par with ranges reported in recent literature on classifying ADs using gradient-boosted and hybrid ensemble classifiers, with overall similar levels of accuracy and F1 where tasks and data are similar [10, 12, 13].

Future research must incorporate evaluation on common publicly available benchmarks (e.g., using the same train/test splits with ADNI, OASIS, or the same Kaggle dataset). It also incorporates the standardization of preprocessing pipelines to reduce variability and measurement of uncertainty (e.g. per-split results and 95% CIs through bootstrapping) and paired-sample tests (e.g., McNemar test to establish accuracy, DeLong test to establish AUC). Calibration and decision-curve analyses to supplement the results are indicated within these limits, we find that hybrid strategies (e.g., Lasso + LightGBM) can produce state-of-the-art dataset competitive performance and practical interpretability in line with the trends of previous work [10, 12, 13].

## 4. CONCLUSIONS

The paper compared conventional, ensemble, and hybrid supervised classifiers in the classification of Alzheimer's disease using tabular clinical data. CatBoost and Lasso + LightGBM (accuracy = 0.96124) were the closest as the strongest point estimate, and XGBoost was considered the third closest (accuracy = 0.96104). All with a strong F1 (0.94 - 0.95). Since we did not report any measures of variance or formal tests of significance, we do not claim to have been statistically better than the other models; instead, the models can be viewed as those that perform best and are statistically equivalent, given the evidence at hand. On a methodological level, the results are congruent with the hypothesis that, with L1-Based selection, features may be denoised and decorrelated, allowing a gradient-boosting learner (LightGBM) to represent nonlinear feature interactions more effectively. Nevertheless, we have seen that the Lasso + LightGBM hybrid cannot be readily interpreted: Lasso produces sparse selections, but the black box model of the final boosted model remains a black box. Future studies will (i) quantify the uncertainty (per-fold results, bootstrap CIs, paired tests such as McNemar/DeLong) to find out whether small metric deltas are statistically significant; (ii) provide explanatory analyses (e.g. SHAP global summaries, local explanations, partial dependence/ICE, and calibration curves) to describe how the output of functional and cognitive measures drives the predictions; (iii) assess blending/stacking on out-of-fold meta-features and probability calibration to increase the diversity among base learners. These criteria suggest that gradient-boosted and hybrid studies are dataset-competitive in AD classification on structured clinical data, and that an additional investigation into uncertainty and explainability is necessary to make comparative or clinical assertions.

## 5. ACKNOWLEDGMENT

The authors owe a great deal to Superior University, Lahore, as it has facilitated this research work by providing the necessary materials and financial support. We appreciate the valuable comments and direction provided by colleagues from the Department of Basic Sciences and the Department of Computer Sciences.

## 6. ETHICAL STATEMENT

The protocols put in place ensured that all research work involving medical data was carried out ethically, employing the best practices. The protocols of this research were reviewed by the Research committee at Superior University Lahore that was chaired by Dr. Muhammad Azam. All materials in this study were confidential and required anonymity.

## 7. CONFLICT OF INTEREST

The authors have no conflict of interest regarding this article.

## 8. REFERENCES

1. D. Jadhav, N. Saraswat, N. Vyawahare, and D. Shirode. Targeting the molecular web of Alzheimer's disease: unveiling pathways for effective pharmacotherapy. *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery* 60(1): 7 (2024).
2. M. Wang, Y. Lin, F. Gu, W. Xing, B. Li, X. Jian, C. Liu, D. Li, Y. Li, T. Wu, and D. Ta. Diagnosis of cognitive and motor disorder levels in stroke patients through explainable machine learning based on MRI. *Medical Physics* 51(3):1763-1774 (2024).
3. F. Öhman, J. Hassenstab, D. Berron, M. Schöll, and K. Papp. Current Advances in Digital Cognitive Assessment for Preclinical Alzheimer's Disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 13(1): 1-19 (2021).
4. P. El Kafrawy, H. Fathi, M. Qaraad, A.K. Kelany, and X. Chen. An efficient SVM-based feature selection model for cancer classification using high-dimensional microarray data. *IEEE Access* 9: 155353-69 (2021).
5. Q.U. Hamza, M.A. Baloch, M.A. Rajwana, A. Raza, and Z.U. Zia. Hybrid ensemble learning approaches for high-accuracy dementia detection: integrating deep learning models. *Kashf Journal of Multidisciplinary Research* 2(05): 66-83 (2025).
6. N. Rane, S.P. Choudhary, and J. Rane. Ensemble Deep Learning and Machine Learning: Applications, Opportunities, Challenges, and Future Directions. *Studies in Medical and Health Science* 1(2): 18-41 (2024).
7. H.T. Hoc, R. Silhavy, Z. Prokopova, and P. Silhavy. Comparing stacking ensemble and deep learning for software project effort estimation. *IEEE Access* 11: 60590-60604 (2023).
8. M.J. Iqbal, Z. Javed, H. Sadia, I.A. Qureshi, A. Irshad, R. Ahmed, K. Malik, S. Raza, A. Abbas, R. Pezzani, and J. Sharifi-Rad. Clinical applications of artificial intelligence and machine learning in cancer diagnosis: Looking into the future. *Cancer Cell International* 21(1): 270-280 (2021).
9. S. Asif, Y. Wenhui, S. Ur-Rehman, Q. Ul-Ain, K. Amjad, Y. Yueyang, S. Jinhai, and M. Awais. Advancements and prospects of machine learning in medical diagnostics: unveiling the future of diagnostic precision. *Archives of Computational Methods in Engineering* 32(2): 853-883 (2024).
10. Y. Jin, Z. Ren, W. Wang, Y. Zhang, L. Zhou, X. Yao, and T. Wu. Classification of Alzheimer's disease using robust TabNet neural networks on genetic data. *Mathematical Biosciences and Engineering MBE* 20(5): 8358-8374 (2023).
11. M. Chakraborty, N. Naoal, S. Momen, and N. Mohammed. ANALYZE-AD: A Comparative Analysis of Novel AI Approaches for Early Alzheimer's Detection. *Array* 22: 100352 (2024).
12. A.S. Alatrany, W. Khan, A. Hussain, H. Kolivand, and D. Al-Jumeily. An Explainable Machine Learning Approach for Alzheimer's Disease Classification. *Scientific Reports* 14(1): 2637-2654 (2024).
13. E. Mahamud, M. Assaduzzaman, J. Islam, N. Fahad, M.J. Hossen, and T.T. Ramanathan. Enhancing Alzheimer's disease detection: An explainable machine learning approach with ensemble techniques. *Intelligence-Based Medicine* 11(11): 100240 (2025).
14. T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona. A survey on missing data in machine learning. *Journal of Big Data* 8(1): 140 (2021).
15. M.K. Dahouda and I. Joe. A deep-learning embedding technique for categorical features encoding. *IEEE Access* 9: 114381-114391 (2021).
16. V. Sharma. A study on data scaling methods for machine learning. *International Journal for Global Academic & Scientific Research* 1(1): 31-42 (2022).
17. R.K. Halder, M.N. Uddin, M.A. Uddin, S. Aryal, and A. Khraisat. Enhancing K-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications. *Journal of Big Data* 11(1): 113-125 (2024).
18. A.A. Amer, S.D. Ravana, and R.A. Habeeb. Effective k-nearest neighbor models for data classification enhancement. *Journal of Big Data* 12(1): 86 (2025).
19. C.K. Reddy, P.A. Reddy, P.S. Reddy, M. Shuaib, S. Alam, S. Ahmad, and A. Rajaram. Twined ensemble framework for network security: integrating Random Forest, AdaBoost, and Gradient Boosting for enhanced intrusion detection. *Discover Internet of Things* 5(1): 107 (2025).
20. H. Şevgin. A comparative study of ensemble methods in the field of education: Bagging and boosting algorithms. *International Journal of Assessment Tools in Education* 10(3): 544-562 (2023).
21. J.C. Timoneda. Estimating group fixed effects in panel data with a binary dependent variable: How

the LPM outperforms logistic Regression in rare events data. *Social Science Research* 93(1): 102486 (2021).

22. A. Demircioğlu. Applying oversampling before cross-validation will lead to high bias in radiomics. *Scientific Reports* 14(1): 11563 (2024).
23. N.S. Nafis and S. Awang. An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification. *IEEE Access* 9: 52177-52192 (2021).
24. T. Mahmood, A. Rehman, T. Saba, T.J. Alahmadi, M. Tufail, S.A. Bahaj, and Z. Ahmad. Enhancing Prognosis of Coronary Artery Disease: A Novel Dual-Class Boosted Decision Trees Strategy for Robust Optimization. *IEEE Access* 12: 107119-107143 (2024).
25. R. Vishraj, S. Gupta, and S. Singh. Evaluation of feature selection methods utilizing random forest and logistic regression for lung tissue categorization using HRCT images. *Expert Systems* 40(8): e13320 (2023).
26. K.A. Ahmed, I. Humaira, A.R. Khan, M.S. Hasan, M. Islam, A. Roy, M. Karim, M. Uddin, A. Mohammad, and M.D. Xames. Advancing breast cancer prediction: Comparative analysis of ML models and deep learning-based multi-model ensembles on original and synthetic datasets. *PLOS One* 20(6): e0326221 (2025).
27. I. Malashin, V. Tynchenko, A. Gantimurov, V. Nelyub, and A. Borodulin. Boosting-based machine learning applications in polymer science: A review. *Polymers* 17(4): 499 (2025).
28. H. Alshamlan, A. Alwassel, A. Banafa, and L Alsaleem. Improving Alzheimer's Disease Prediction with different Machine Learning Approaches and Feature Selection Techniques. *Diagnostics* 14(19): 2237 (2024).
29. R.A. Gad and A. Abdelhafeez. Alzheimer's Disease Prediction using Hybrid Machine Learning Techniques. *SciNexuses* 1: 174-83 (2024).