A Comparative Analysis of Machine Learning Models for Alzheimer's Disease Screening

1st Bokhtiar Mehedy
Blekinge Institute of Technology
Karlskrona, Sweden

2nd Kidus Mikael Birhanu
Blekinge Institute of Technology
Karlskrona, Sweden

***Abstract*—This study develops a machine learning approach for binary classification of Alzheimer's disease using a dataset of 2,149 instances and 32 features. A systematic workflow was followed, including feature selection, model evaluation, and interpretability analysis. Models like SVM, Random Forest, XGBoost, and CatBoost, along with a stacked ensemble, were compared. CatBoost slightly outperformed other models; however, Random Forest was selected due to its superior interpretability. The results demonstrate the potential of explainable machine learning in healthcare applications.**

I. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the leading cause of dementia globally, accounting for 60-70% of cases [1]. It progressively damages brain cells, resulting in memory loss, cognitive decline, loss of basic abilities, and ultimately, death. Treating AD currently costs $1 trillion annually, with 152 million cases projected by 2050, predominantly in low- and middle-income countries. Despite 10 million new dementia cases reported yearly, most remain undiagnosed, with only 20-50% documented in high-income countries and as low as 10% in low-income regions [2].

There are no effective treatments for most AD patients, and while its primary causes remain unknown (except for a small subset of familial cases linked to genetic mutations) early diagnosis can delay disease progression and improve quality of life. Current diagnostic methods include non-invasive approaches such as cognitive tests, behavioral evaluations, and imaging techniques (CT, MRI, and PET scans), as well as invasive options like cerebrospinal fluid analysis or biomarker blood tests [3]. However, these processes are often complex and costly.

Our objective is to analyze collected data and develop a machine learning model capable of screening potential patients through a few simple online questions. For this study, we used a publicly available Alzheimer's disease dataset from Kaggle [4], which contains structured clinical data for Alzheimer's diagnosis and prediction. This initial screening would identify potential cases for further clinical evaluation to confirm an Alzheimer's diagnosis. We prioritized explainability and interpretability during model selection to build trust among domain experts, such as doctors, and enhance patient confidence.

In our model selection process, we performed detailed data analysis, applying techniques like scaling, encoding, and feature selection as required. We evaluated six models and, through rigorous testing, shortlisted the two best-performing ones for final evaluation. Experiments with different data distribution combinations allowed us to refine our findings with thorough analysis and clear motivation. Throughout the process, we integrated concepts learned during the course to ensure a structured and methodical workflow.

This report is organized as follows: Section 2 outlines the methodology, detailing the flow of work conducted, including data analysis, model training, evaluation, and selection. Section 3 presents our findings with a detailed discussion. Section 4 summarizes the study by highlighting key results, and Section 5 outlines the contributions of the authors.

II. Methodology

A. Data Understanding and Pre-processing

The dataset, obtained from Kaggle, consists of 2,149 observations with 32 features, of which 15 are numerical and 17 are categorical. The target variable is binary, where 0 represents non-Alzheimer's cases, and 1 represents Alzheimer's cases.

All values in the dataset are non-null and numerical, with no missing values or duplicate records. The class distribution is moderately imbalanced, with 35.4% of instances representing Alzheimer's cases and 64.6% representing non-Alzheimer's cases.

1) Data balancig: Due to the moderate imbalance in the data, we needed to balance it to avoid overfitting. When experimenting with oversampling using SMOTE, we observed poor results due to higher misclassification rates for Alzheimer's cases compared to non-Alzheimer cases. To prevent this issue from being further amplified, we opted for undersampling, which provided more stable and reliable results for our dataset.

Initially, we trained and evaluated the model without balancing the data, achieving high metrics, such as 97% accuracy. However, we observed that the results were slightly biased towards the non-Alzheimer's class, leading to overfitting. As a result, the model's performance in detecting Alzheimer's cases (class 1) was lower than expected. To address this, we repeated the entire process with balanced data, ensuring a more reliable and fair evaluation of the model's predictive capabilities.

2) Data Encoding: Among the 17 categorical features, two (Ethnicity and Education level) had four classes. To handle categorical variables, one-hot encoding was applied to these features, while all other categorical features were binary and required no further transformation.

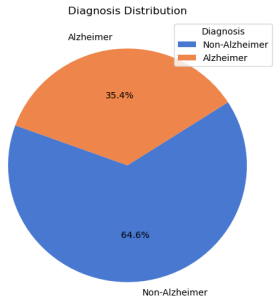

Fig. 1. Class distribution

3) Feature Selection: To identify the most relevant features, multiple feature selection techniques were employed:

- Statistical and correlation-based methods: Spearman rank correlation, mutual information.
- Dimensionality reduction techniques: Principal Component Analysis (PCA).
- Model-based selection methods: Recursive Feature Elimination (RFE)

These feature selection methods were applied, each yielding different top features. To ensure robustness, we ranked features based on their frequency of selection across methods and retained those appearing in at least four techniques(15 features). Notably, five features (FunctionalAssessment, ADL, MemoryComplaints, MMSE, and BehavioralProblems) consistently ranked in the top five across all selection methods, confirming their strong predictive importance.

B. Model Selection

Initially, six machine learning models were considered for Alzheimer's disease prediction:

- Decision Tree
- Random Forest
- CatBoost
- XGBoost
- AdaBoost
- Support Vector Machine

The selected machine learning models were chosen for their ability to handle structured medical data and capture complex patterns relevant to Alzheimer's disease prediction. Decision Tree (DT) provides an interpretable baseline but is prone to overfitting. Random Forest (RF) improves stability by averaging multiple trees, making it more robust. CatBoost efficiently handles categorical features and is optimized for structured datasets, while XGBoost enhances predictive performance through boosting, making it effective for imbalanced data. Finally, Support Vector Machine (SVM) is included for its strong performance in high-dimensional, non-linear classification tasks.

C. Build Models

The models were trained using an 80-20 data split, and their predictive performance was evaluated across various metrics.

Model	Accuracy	Precision	Recall	F1-Score
CatBoost	0.938	0.938	0.938	0.937
Random Forest (RF)	0.934	0.935	0.934	0.934
XGBoost	0.924	0.924	0.924	0.924
Decision Tree (DT)	0.882	0.882	0.882	0.882
AdaBoost	0.885	0.891	0.884	0.884
(SVM)	0.875	0.875	0.875	0.875

TABLE I
PERFORMANCE METRICS OF DIFFERENT MODELS

CatBoost (0.938) emerged as the best-performing model, followed closely by Random Forest (0.934) and XGBoost (0.924), all showing strong predictive capabilities. Decision Tree, AdaBoost, and SVM performed moderately (0.88), with SVM having the lowest F1-score (0.875). Given the results, we selected the top three models for further evaluation.

D. Cross Validation

Model	Accuracy	Precision	Recall	F1-Score
CatBoost	0.947	0.948	0.947	0.946
Random Forest (RF)	0.940	0.942	0.940	0.940
XGBoost	0.939	0.941	0.939	0.939

TABLE II
FINAL CROSS-VALIDATION RESULTS (AVERAGE OVER 10 FOLDS)

After stratified cross-validation, CatBoost and Random Forest remained the top performers, while XGBoost lagged slightly. Based on this, we proceeded with hyperparameter tuning for CatBoost and Random Forest, dropping XGBoost from further evaluation.

E. Hyper Parameter Tuning

Model	Accuracy	Precision	Recall	F1-Score
CatBoost	0.941	0.941	0.941	0.941
Random Forest (RF)	0.938	0.938	0.938	0.937

TABLE III
FINAL HYPER PARAMETER TUNING

Interestingly, hyper-parameter tuning showed slightly reduced performance compared to cross-validation. Investigating CatBoost's folds, we found some achieving 98%, likely inflating the average cross-validation score and overestimating the model's true performance.

F. Stacking

To enhance prediction performance, a stacking ensemble was constructed using the top two models (Random Forest & CatBoost). The outputs of these models were combined to create a meta-model for better generalization.

Model	Accuracy	Precision	Recall	F1-Score
CatBoost (RF)	0.940	0.942	0.940	0.940
Random Forest	0.924	0.925	0.924	0.924
Stacking	**0.941**	**0.941**	**0.941**	**0.941**

TABLE IV
STACKING CLASSIFIER AND INDIVIDUAL MODELS

Since the stacking model did not significantly outperform the individual models, we proceeded with the Friedman test to determine if there was a statistically significant difference between the three models.

G. Friedman Test

The Friedman test was conducted to statistically validate whether the observed performance differences among the models were significant. Given the small performance gap between CatBoost, Random Forest, and the Stacking model, it was essential to ensure that these differences were not due to random variations in the dataset. This test allowed us to make a more data-driven and reliable decision when selecting the final model.

Rank	Model	AVG Rank
1	CatBoost	1.40
2	Random Forest	2.00
3	Stacking	2.60

TABLE V
FRIEDMAN TEST

After performing the Friedman test, we observed that CatBoost outperformed both the Stacking model and Random Forest. Since stacking did not significantly improve performance as expected, we proceeded with the final model selection between Random Forest and CatBoost.

H. Final Model Selection

Before selecting our final model, we evaluated the ROC curve and AUC score for both Random Forest (RF) and CatBoost. A screenshot is provided, showing that CatBoost achieved an AUC of 0.95 and an ROC score of 0.948.

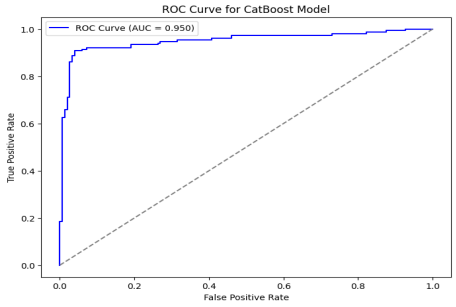

Fig. 2. ROC Curve for

Despite CatBoost outperforming Random Forest in terms of overall performance, we prioritized interpretability over slight performance gains due to the model's intended application in the medical sector, where transparency and explainability are crucial.

1) Interpretability: For Random Forest, we conducted an interpretability analysis, which confirmed that the same five key features (FunctionalAssessment, ADL, MemoryComplaints, MMSE, and BehavioralProblems) played the most significant role in classification. This further reinforced the model's transparency and reliability in decision-making.

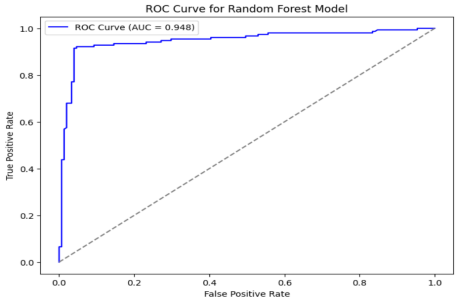

Fig. 3. ROC Curve for Random Forest

Since this model is intended for use in the medical sector, interpretability is crucial to ensure that clinicians and healthcare professionals can understand and trust the predictions, making the model more practical for real-world implementation.Interpretability methods for tree-based models, such as feature importance ranking, have been widely studied [5]. A screenshot of the interpretability analysis is provided.

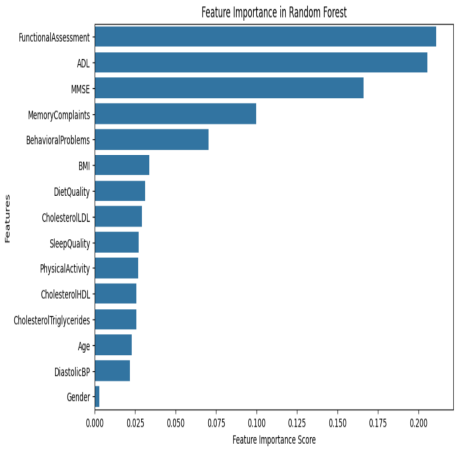

Fig. 4. RF Feature importance

2) Explainability: We performed explainability analysis only for CatBoost, as Random Forest has built-in interpretability, making additional explainability analysis unnecessary. Using SHAP, we identified that only a few features significantly influenced the model's classification, with the same five key features standing out. SHAP provides a unified approach to explaining tree-based models, enhancing transparency in AI-driven decision-making [5]. A screenshot of the SHAP analysis is provided(Fig. 5).

3) Model Selection Decision: While CatBoost achieved slightly better overall performance (Accuracy: 0.941, Precision: 0.941, Recall: 0.941, F1-score: 0.941) compared to Random Forest (Accuracy: 0.938, Precision: 0.938, Recall: 0.938, F1-score: 0.937), the difference was marginal.

However, when specifically evaluating the models' ability to detect Alzheimer's patients (class 1), CatBoost demonstrated slightly better sensitivity than Random Forest, as shown in Fig. 6. The confusion matrices indicate that CatBoost correctly

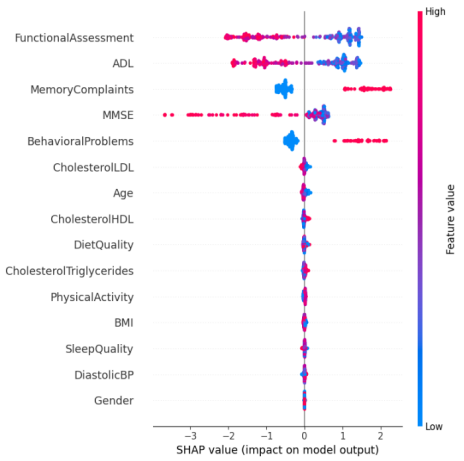

Fig. 5. CatBoost SHAP

identified one additional Alzheimer's case, with a True Positive (TP) count of 142 and False Negatives (FN) of 11, whereas Random Forest had a TP count of 141 and FN of 12. This minor difference suggests that CatBoost had a slightly better recall in detecting Alzheimer's cases, reducing the risk of misclassification for patients who may require medical attention.

Despite CatBoost's marginally better performance, we selected Random Forest as the final model due to its interpretability, which is crucial in the medical field for trust and decision-making. This choice ensures that clinicians and researchers can understand how the model arrives at its predictions, making it more suitable for real-world implementation in healthcare, where transparency and explainability are critical factors.

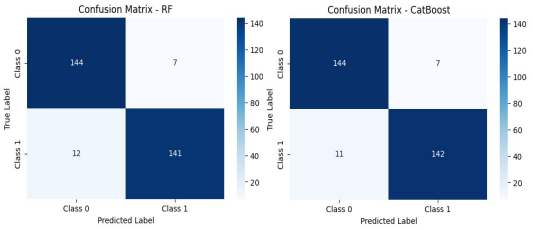

Fig. 6. Confusion matrix (RF and Catboost)

I. Gender-Based Performance Analysis

After dividing the dataset by gender, we evaluated the Random Forest (RF) model separately for male and female groups. The results reveal a notable difference in model performance between the two groups.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest (RF)	0.895	0.892	0.899	0.894

TABLE VI
MALE

Model	Accuracy	Precision	Recall	F1-Score
Random Forest (RF)	0.941	0.941	0.941	0.941

TABLE VII
FEMALE

These results suggest that the model is more effective in predicting Alzheimer's disease in females than in males. The higher performance in females may indicate differences in feature distribution, symptom patterns, or data representation between the two groups. This warrants further investigation to determine whether To improve fairness and generalizability, further analysis and potential adjustments in feature selection, data balancing, or model training may be necessary to ensure equitable performance across both genders.

J. Feature-Based Model Evaluation

The results show that using only the top 5 features achieved nearly identical or slightly improved performance compared to using all 15 features. This indicates that the top 5 features alone are sufficient for accurate classification, rendering the additional features redundant in our case.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest (RF)	0.937	0.938	0.938	0.937

TABLE VIII
15

Model	Accuracy	Precision	Recall	F1-Score
Random Forest (RF)	0.940	0.941	0.941	0.941

TABLE IX
TOP 5

III. Results Analysis and Discussion

A. Top 3 Model Selection Based on Performance

Initially, six machine learning models were evaluated for Alzheimer's disease prediction. Based on their performance and robustness from Table 1, the top three models were selected for further analysis: CatBoost, Random Forest, and XGBoost. These models were chosen because they consistently outperformed the other candidates in terms of accuracy, recall, and overall predictive power, making them reliable for Alzheimer's classification. Additionally, ensemble-based models (CatBoost, Random Forest, and XGBoost) are known for their ability to handle structured medical data, mitigate overfitting, and effectively manage feature importance. CatBoost emerged as the best performer, followed closely by Random Forest and XGBoost, justifying their selection for further evaluation.

B. Cross-Validation

To ensure robustness and generalizability, we performed 10-fold stratified cross-validation. The results, as shown in Table II, indicated that CatBoost and Random Forest remained the top-performing models, while XGBoost lagged slightly

behind. Given the marginal difference in performance and the need to streamline model selection, we decided to drop XGBoost and proceed with hyperparameter tuning for CatBoost and Random Forest to further optimize their performance.

C. Improvement After Hyperparameter Tuning

After hyperparameter tuning, both CatBoost and Random Forest showed slight performance improvements, with CatBoost outperforming Random Forest overall. However, these results were slightly lower than the cross-validation scores. Further analysis revealed that some CatBoost cross-validation folds achieved exceptionally high accuracy (98%), likely inflating the average and overestimating the model's true effectiveness.

D. Stacking to Further Enhance Performance

The stacking model (Random Forest + CatBoost) showed a slight decrease in performance compared to CatBoost (RF), with accuracy, precision, recall, and F1-score all lower. This suggests that, in this case, stacking did not significantly improve performance and may have slightly reduced it compared to using CatBoost alone.

E. Friedman Test for Significant Differences

The Friedman test was performed to statistically compare the three models. The results showed that CatBoost ranked the highest, followed by Random Forest and Stacking, confirming that Stacking did not provide a meaningful advantage over the individual models.

F. Final Model Selection: Prioritizin Interpretability

Despite CatBoost achieving the best performance, we prioritized interpretability over slight performance gains due to the medical application of our model. Medical professionals need to understand and trust the model's predictions, making interpretability crucial for real-world implementation.

- CatBoost (Accuracy: 0.941, Precision: 0.941, Recall: 0.941, F1-score: 0.941)
- Random Forest (Accuracy: 0.938, Precision: 0.938, Recall: 0.938, F1-score: 0.937)

While the difference in performance was marginal, Random Forest was selected as the final model due to its built-in interpretability, making it more suitable for healthcare applications. This ensures that clinicians and researchers can trace back the decision-making process, increasing trust and usability in medical diagnostics.

IV. Conclusions

This study aimed to develop a machine learning model for early Alzheimer's disease (AD) detection through simple online questions, prioritizing interpretability to facilitate adoption in medical practice. After rigorous data preprocessing, feature selection, and model evaluation, we found that CatBoost, Random Forest, and XGBoost were the top-performing models. Despite CatBoost's superior performance, Random Forest was ultimately selected as the final model due to its interpretability, a critical factor in medical diagnostics where understanding the model's decision-making process is essential.

Our results highlighted the importance of balancing performance with model transparency, as clinicians and healthcare professionals require clear insights into the factors influencing predictions. While CatBoost showed marginally higher accuracy, the Random Forest model's built-in feature importance and ease of explanation make it more suitable for practical implementation in healthcare.

The study also revealed some notable findings, such as the better performance of the Random Forest model in predicting AD in females compared to males, suggesting that further investigation into gender-based data discrepancies is needed. Additionally, our analysis demonstrated that a reduced set of five top features could achieve similar or even slightly better performance than the full set, suggesting that simpler models could be more effective and computationally efficient.

In conclusion, the selected Random Forest model, with its emphasis on interpretability, provides a solid foundation for screening potential Alzheimer's cases, which can be followed by further clinical evaluation. The model's strong performance in this task indicates that machine learning can be a valuable tool in addressing the growing challenge of Alzheimer's diagnosis, especially in resource-limited settings.

Acknowledgment

We sincerely thank Veselka Boeva and Shahrooz Abghari for their valuable guidance, feedback, and support throughout this project. Their insights greatly contributed to refining our methodology and enhancing our findings.

References

[1] World Health Organization, "Dementia Fact Sheet," [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia.

[2] Alzheimer's Disease International, "Dementia Statistics," [Online]. Available: https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/.

[3] National Institute on Aging, "Alzheimer's Disease Fact Sheet," [Online]. Available: https://www.nia.nih.gov/health/alzheimers-and-dementia/alzheimers-disease-fact-sheet.

[4] R. El Kharoua, "Alzheimer's Disease Dataset," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset/data.

[5] J. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," arXiv preprint, [Online]. Available: https://arxiv.org/pdf/1811.10154.