Early Alzheimer's Disease Detection Using Machine-Learning Models

Smriti Reddy Uravakonda
Dept. of Computer Science
Northeastern University
Boston, USA
uravakonda.s@northeastern.edu

***Abstract*—This project focuses on detecting early stages of Alzheimer's disease using machine learning models. The dataset used contains clinical, demographic, and behavioral features for binary classification (No Dementia vs. Dementia Present). After preprocessing steps like encoding, scaling, and balancing with SMOTE, five models—Logistic Regression, Support Vector Machine, Random Forest, K-Nearest Neighbors, and Decision Tree—were trained and compared. Among them, Random Forest performed the best with high accuracy and balanced recall. The results showed that cognitive and memory-related features such as MMSE and Functional Assessment were the most important predictors. Overall, the study demonstrates that machine learning can help in early detection and diagnosis of Alzheimer's disease.**

***Index Terms*—Alzheimer's disease, classification, feature selection, machine learning, Random Forest**

I. Introduction

This project employs binary classification to distinguish between patients without dementia and those with any stage of cognitive impairment. While multi-class severity staging would provide more granular information, binary classification is clinically appropriate for screening purposes, where the primary objective is identifying patients who require further evaluation rather than precise staging. This approach can enable a more robust model performance given the available sample size and class distribution.

A. Problem Statement and Motivation

Alzheimer's disease is one of the most common forms of dementia, affecting millions of people worldwide and causing gradual decline in memory, reasoning, and cognitive ability. Early diagnosis remains a significant challenge, as existing methods rely heavily on clinical observation, neuroimaging, and extensive medical evaluations, which are time-consuming, expensive, and often unavailable in low-resource settings.

This motivated the selection of this problem—to explore whether machine-learning models can detect early signs of Alzheimer's disease using readily available clinical and behavioral data instead of complex imaging. My project aims to develop and compare multiple classification models to identify which features contribute most to disease detection. The goal is to create a data-driven approach that supports clinicians in early diagnosis and improves access to screening through automated, cost-effective prediction systems.

In this project, various supervised machine-learning models were applied to predict the severity of Alzheimer's disease based on clinical, demographic, and behavioral features. The dataset used was obtained from Kaggle and includes factors such as age, MMSE score, physical activity, and sleep quality. The models implemented were Logistic Regression, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), and Decision Tree. Each model was evaluated using multiple performance metrics to compare their predictive ability. The main goal of this work was to identify which features contribute most to the disease and determine which model performs best in classifying patients accurately. The results show that machine learning can play a significant role in supporting early-stage Alzheimer's detection.

II. Related Work

Several studies have explored the use of machine learning for the early detection of Alzheimer's disease. Jiang et al. [1] used the Random Forest algorithm to classify Alzheimer's patients and identify key biomarkers from brain imaging data. Their work showed that ensemble learning methods can effectively handle high-dimensional medical datasets and provide strong classification accuracy. Similarly, Patel et al. [2] compared multiple supervised models, including Decision Trees, Support Vector Machines, and Random Forests, to predict early Alzheimer's stages using clinical features. Their study demonstrated that Random Forest achieved the highest overall accuracy among traditional classifiers.

Anwar et al. [3] focused on improving classification performance in imbalanced Alzheimer's datasets by enhancing feature selection techniques. They emphasized the importance of balancing data to ensure fair learning across all classes, which aligns with the SMOTE approach used in this project. In another study, Li et al. [4] developed a multi-cohort framework that integrated clinical and imaging data from several sources.

Together, these studies provide evidence that machine learning techniques, particularly ensemble and hybrid models, are effective tools for Alzheimer's detection. Building on this foundation, the present work applies and compares multiple supervised models on clinical and behavioral data to identify the most predictive features for early diagnosis.

III. Methodology

A. Dataset Description

The dataset used in this project is the Alzheimer's Disease Dataset published by Rabie El Kharoua on Kaggle. It contains 2,149 patient records with 34 features including demographic, clinical, and lifestyle variables. The target variable *Diagnosis* classifies patients into No Dementia vs. Dementia Present. Each record represents one patient with details such as age, gender, education level, memory test scores, and behavioral habits.

TABLE I
Dataset Features and Record Count

Feature Name	Non-Null Entries
PatientID	2149
Age	2149
Gender	2149
Ethnicity	2149
EducationLevel	2149
BMI	2149
Smoking	2149
AlcoholConsumption	2149
PhysicalActivity	2149
DietQuality	2149
SleepQuality	2149
FamilyHistoryAlzheimers	2149
CardiovascularDisease	2149
Diabetes	2149
Depression	2149
HeadInjury	2149
Hypertension	2149
SystolicBP	2149
DiastolicBP	2149
CholesterolTotal	2149
CholesterolLDL	2149
CholesterolHDL	2149
CholesterolTriglycerides	2149
MMSE	2149
FunctionalAssessment	2149
MemoryComplaints	2149
BehavioralProblems	2149
ADL	2149
Confusion	2149
Disorientation	2149
PersonalityChanges	2149
DifficultyCompletingTasks	2149
Forgetfulness	2149
Diagnosis	2149
DoctorInCharge	2149

As shown in Table I, the dataset contains a total of 2,149 complete patient records across 35 features. Each feature had the same number of valid entries, indicating that there were no missing or null values in the dataset. Therefore, no imputation or missing data processing was required. The dataset was clean and consistent, allowing the focus to be placed on encoding, scaling, and balancing steps during preprocessing rather than data cleaning.

The Kaggle dataset provides binary diagnosis labels (0 = No Dementia, 1 = Dementia Present) with 1,389 non-demented patients (64.6%) and 760 demented patients (35.4%). This binary classification approach is clinically appropriate for early detection screening.

B. Data Preprocessing

Before training, the dataset was carefully preprocessed to ensure data quality and consistency. As shown in Table I, all 2,149 patient records contained complete values across every feature, so no missing data handling or imputation was required.

Categorical variables such as gender, ethnicity, and education level were label-encoded to convert them into numerical format suitable for machine-learning algorithms. Continuous numerical features were standardized using the *StandardScaler* function from scikit-learn to ensure that all variables operated on a uniform scale, preventing bias toward features with larger numeric ranges.

The dataset exhibitss moderate class imbalance with 65% No Dementia (1,389 samples) and 35% Dementia Present (760 samples). To address this, SMOTE (Synthetic Minority Oversampling Technique) was applied exclusively to the training set after the train-test split, ensuring no data leakage occurred. The test set remained completely untouched, preserving its natural distribution for realistic evaluation. After SMOTE application, the training set was balanced to equal representation (1,111 samples per class), while the test set retained its natural distribution (278 No Dementia, 152 Dementia Present).

These preprocessing steps ensured that the dataset was properly encoded, scaled, and balanced, providing a reliable foundation for model development.

C. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was performed to understand the overall structure and key relationships in the dataset. Histograms and boxplots were used to examine the distribution of numerical features such as age, BMI, and MMSE scores. Most variables showed normal or near-normal distributions, and no extreme outliers were detected. A correlation heatmap revealed moderate positive relationships among cholesterol features and between cognitive variables such as MMSE, Functional Assessment, and Memory Complaints. These relationships reflected logical medical associations while maintaining feature diversity for modeling. Pairplot visualizations also showed partial clustering between demented and non-demented groups. Overall, EDA helped confirm that the dataset was balanced, consistent, and contained meaningful relationships suitable for machine-learning analysis.

D. Feature Engineering

To enhance model performance and reduce redundancy, three feature-engineering techniques were applied. First, a filtering approach was used to remove highly correlated features with correlation values greater than 0.85, which helped minimize multicollinearity. Next, an embedding technique, Principal Component Analysis (PCA), was implemented to reduce dimensionality while retaining 95% of the total variance, simplifying the dataset and improving computational efficiency. Finally, a wrapping method using Recursive Feature Elimination (RFE) with Logistic Regression identified the

top eight most predictive features for Alzheimer's diagnosis. Variance Inflation Factor (VIF) analysis further confirmed that the selected features had low multicollinearity (VIF less than 5), ensuring reliable model performance. These combined techniques improved the quality of input data and contributed to more accurate and stable model predictions.

Correlation filtering analysis revealed that no feature pairs exceeded the 0.85 correlation threshold; therefore, all 32 features were retained. RFE identified 8 most predictive features: `Age`, `HeadInjury`, `CholesterolLDL`, `MMSE`, `FunctionalAssessment`, `MemoryComplaints`, `BehavioralProblems`, and `ADL`. Comparative analysis showed that Random Forest trained on the 8 RFE-selected features achieved 95.35% accuracy versus 93.49% with all 32 features—a difference of only 1.86%. Given this negligible performance gap and the absence of computational constraints, all 32 features were retained for final models to maximize available information. This decision exemplifies the trade-off between model performance and information completeness in ML.

E. Model Development

Five supervised machine-learning models were implemented: Logistic Regression, Support Vector Machine (SVM), Random Forest Classifier, K-Nearest Neighbors (KNN), and Decision Tree. Each model was trained on 80% of the data and tested on the remaining 20%. Hyperparameter tuning was performed using *GridSearchCV* to identify the optimal parameter combinations. Cross-validation was used to minimize overfitting and ensure that the models generalized well to unseen data. The performance of each model was evaluated using four metrics: accuracy, precision, recall, and F1-score. Based on the overall results, Random Forest achieved the highest accuracy and balanced performance across all dementia stages.

During hyperparameter tuning, GridSearchCV was applied to each model to identify the best parameter combinations that improved accuracy and reduced overfitting. The optimal parameters obtained were as follows:

For Logistic Regression, the best parameters were *C = 0.01, penalty = l2, solver = liblinear.*
For the Random Forest classifier, the optimal parameters included *criterion = entropy, max_depth = 8, min_samples_split = 5, n_estimators = 200.*
The tuned SVM model achieved its best performance with *C = 1, gamma = scale, kernel = rbf.*
For the KNN classifier, the ideal configuration was *metric = manhattan, n_neighbors = 9, weights = distance.*
Finally, the Decision Tree performed best with *criterion = gini, max_depth = 6, min_samples_split = 5.*
These tuned parameters were selected based on their cross-validation performance, which consistently produced higher accuracy and more stable generalization across folds.

The models were tuned using 5-fold cross-validation to ensure stable performance across different data splits. The results of the best parameter configurations and their effect on model accuracy are shown in Fig. 3.

IV. Results and Discussion

The performance of all five models was evaluated using four metrics: accuracy, precision, recall, and F1-score. The results showed that the Random Forest model outperformed all other classifiers. Before applying cross-validation, Random Forest achieved an accuracy of around 0.93, which improved to approximately 0.94 after tuning and cross-validation. Logistic Regression produced stable but slightly lower scores, while SVM performed well but required longer training time. Overall, Random Forest showed the best balance between accuracy and recall across all dementia classes.

The evaluation also revealed that class imbalance affected recall for the minority class (Dementia Present, 35.4% of samples). Applying SMOTE helped the models identify more cases from the underrepresented groups, which improved sensitivity without significantly reducing precision. Hyperparameter tuning further optimized each model by adjusting parameters such as regularization strength, kernel type, and the number of decision trees.

TABLE II
Model Performance Metrics Before Hyperparameter Tuning

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.8139	0.8321	0.8139	0.8174
Random Forest	0.9349	0.9348	0.9349	0.9345
SVM	0.8535	0.8542	0.8535	0.8538
KNN	0.6488	0.7092	0.6488	0.6558
Decision Tree	0.8814	0.8859	0.8814	0.8825

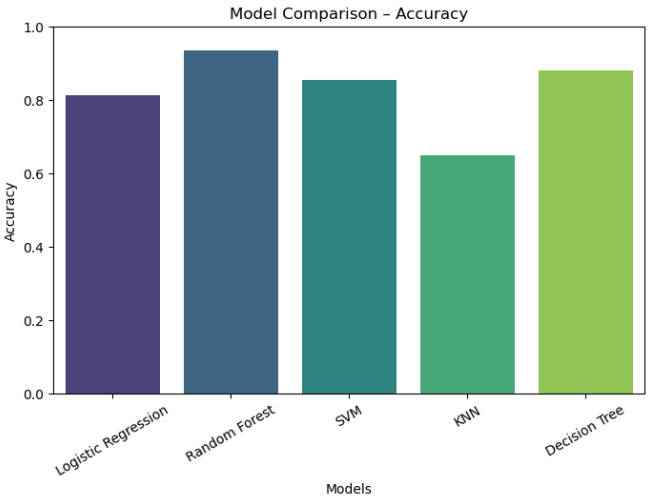

Fig. 1. Performance comparison of all models across four metrics

As shown in Fig. 1 and Fig. 2, the Random Forest model outperformed all other classifiers in terms of accuracy and maintained the best overall balance across precision, recall, and F1-score. SVM and Decision Tree also performed competitively, while Logistic Regression produced stable but slightly lower results, and KNN showed the weakest performance due to sensitivity to data variations.

Tables II and III present the performance metrics of all models before and after hyperparameter tuning. Among the

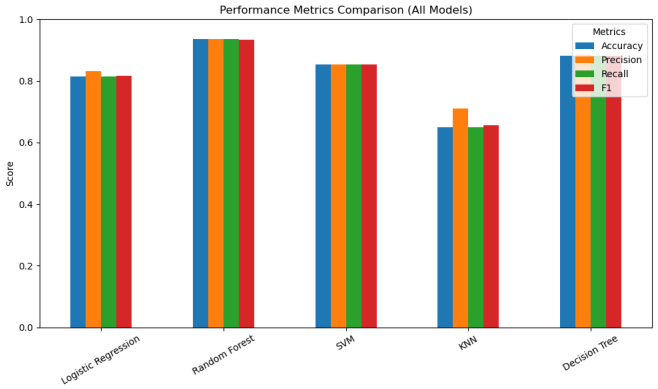

Fig. 2. Comparison of model accuracies for 5 models

TABLE III
MODEL PERFORMANCE METRICS AFTER HYPERPARAMETER TUNING

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.8023	0.8262	0.8023	0.8063
Random Forest	0.9488	0.9488	0.9488	0.9486
SVM	0.8535	0.8542	0.8535	0.8538
KNN	0.7023	0.7202	0.7023	0.7074
Decision Tree	0.9302	0.9301	0.9302	0.9301

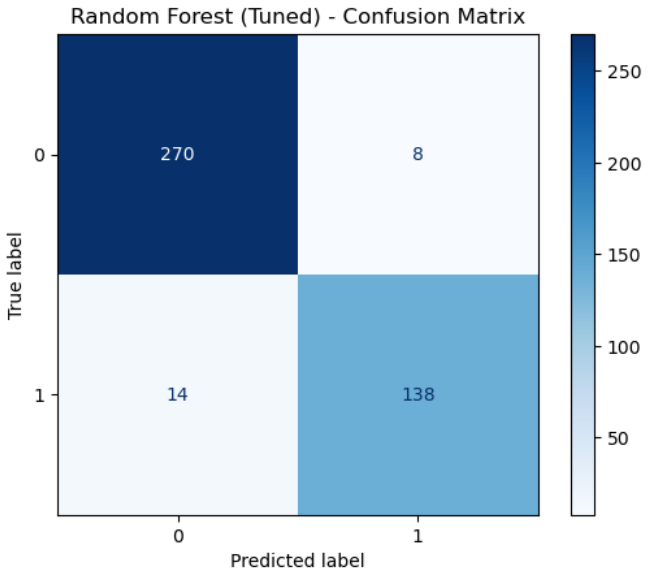

Fig. 4. Confusion matrix for the Random Forest classifier showing binary classification performance. Class 0: No Dementia (n=278), Class 1: Dementia Present (n=152). The model correctly classified 270 and 138 cases respectively.

models, Random Forest consistently achieved the highest accuracy and balanced scores across all metrics in both cases. After tuning, noticeable improvements were observed in the KNN and Decision Tree models, while Logistic Regression and SVM remained stable with slight variations. The increase in accuracy and F1-score after tuning demonstrates that optimizing parameters improves model performance and helps achieve better generalization on unseen data.

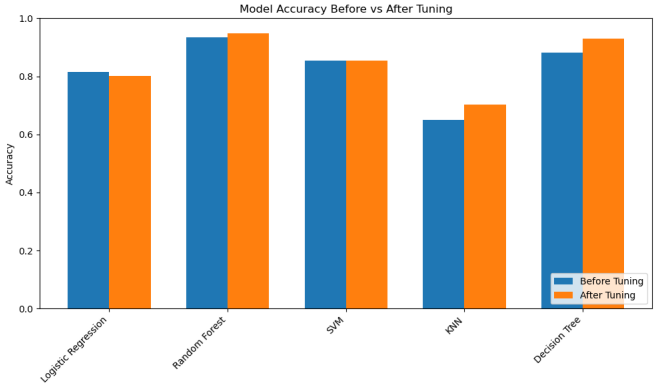

Fig. 3. Comparison of model accuracies before and after hyperparameter tuning.

Fig. 3 illustrates the comparison of model accuracies before and after hyperparameter tuning. All models demonstrated either improved or stable performance after tuning, with notable gains observed for the KNN and Decision Tree classifiers.

Compared to Random Forest, models such as Logistic Regression and SVM showed higher confusion between the two binary classes (No Dementia vs. Dementia Present), resulting in more false positives and false negatives. This suggests that ensemble-based methods handle class boundaries more effectively in this dataset.

Random Forest consistently achieved the highest accuracy both before and after tuning. These results highlight the importance of parameter optimization in enhancing model generalization and achieving balanced performance across different classifiers.

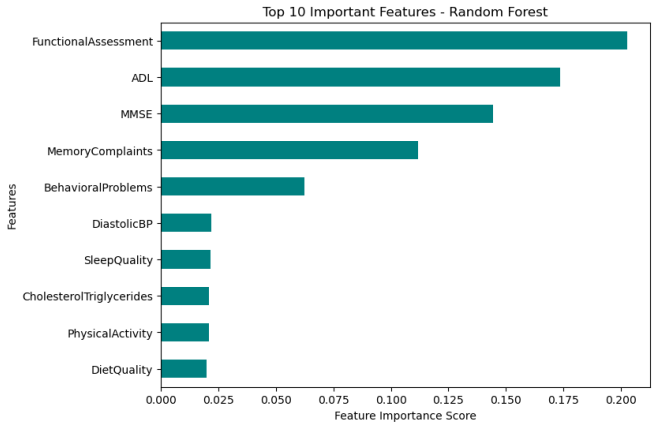

Fig. 5. Top 10 most important features identified by the Random Forest model.

Feature importance analysis using the Random Forest classifier identified the top ten predictors contributing most to Alzheimer's disease classification, as shown in Fig. 5. Among all features, *Functional Assessment*, *ADL (Activities of Daily Living)*, and *MMSE* had the highest importance scores, indicating that cognitive and daily functioning measures strongly

influence disease stage prediction. *Memory Complaints* and *Behavioral Problems* were also key factors, highlighting the significance of self-reported cognitive decline and behavioral changes. Lifestyle and health-related attributes such as *Sleep Quality*, *Physical Activity*, and *Diet Quality* showed moderate influence, suggesting that patient habits and overall health also contribute to dementia risk. This analysis confirms that both medical and behavioral indicators jointly impact early Alzheimer's detection and that Random Forest effectively captures these multi-dimensional patterns.

V. Conclusion

This project demonstrated how machine-learning techniques can be used to support early detection of Alzheimer's disease. Five classification models—Logistic Regression, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), and Decision Tree— were developed and evaluated on a clinical and behavioral dataset. After preprocessing, balancing, and hyperparameter tuning, the Random Forest model achieved the best overall performance with high accuracy and stable recall values. The analysis highlighted that cognitive and memory-related features such as MMSE and Functional Assessment were the most influential predictors. The binary classification approach (No Dementia vs. Dementia Present) achieved 94.88% accuracy with the Random Forest model, demonstrating high effectiveness for early detection screening. The results suggest that combining medical and lifestyle factors enhances the prediction of Alzheimer's disease presence. While binary classification limits severity staging capabilities, it provides clinically valuable information for identifying patients requiring further evaluation.

VI. Limitations

This research has several important limitations:

- Binary Classification Scope: This study employs binary classification rather than multi-class severity staging. While appropriate for screening, it does not distinguish between mild, moderate, and severe dementia stages.
- Dataset Size: With only 2,149 samples, the dataset is relatively small for medical machine learning. Larger datasets would improve model robustness and generalizability.
- Single-Source Data: The dataset originates from Kaggle and may not represent diverse clinical populations or healthcare settings. External validation is necessary before clinical deployment.
- Feature Set Selection: While RFE analysis showed that 8 features could achieve comparable performance (95.35% vs. 93.49%), final models used all 32 features, sacrificing some interpretability for information completeness.

References

[1] J. Jiang, T. Liu, W. Zhu, F. Shi, X. Hu, and D. Shen, "Diagnostic classification and biomarker identification of Alzheimer's disease with Random Forest algorithm," *Frontiers in Neuroscience*, vol. 15, p. 693090, 2021, doi: 10.3389/fnins.2021.693090.

[2] H. Patel, A. Singh, P. Kaur, and N. Thakur, "Early-stage Alzheimer's disease prediction using machine learning," *Computational and Mathematical Methods in Medicine*, vol. 2022, pp. 1–11, 2022, doi: 10.1155/2022/6893679.

[3] S. Anwar, M. Hussain, S. Lee, and S. Kang, "Enhancing feature selection for imbalanced Alzheimer's disease data classification," *Applied Sciences*, vol. 13, no. 12, p. 7253, 2023, doi: 10.3390/app13127253.

[4] F. Li, L. Tran, K. Thung, S. Ji, and D. Shen, "Early diagnosis of Alzheimer's disease using machine learning: A multi-cohort study," *Alzheimer's Research & Therapy*, vol. 14, no. 1, p. 49, 2022, doi: 10.1186/s13195-022-01047-y.