Alzheimer's Disease Detection: A real-data comparison of the accuracy using some machine learning models

1st Quazi Reshoan Yazdi 2nd MD. Iftear Hossain Ratul 3rd MD. Shazidul Islam 4th Kazi Istiak Ahammed
Department of Computer Science
American International University Bangladesh
Dhaka, Bangladesh
21-45485-3@student.aiub.edu 21-45678-3@student.aiub.edu 21-45988-3@student.aiub.edu
20-43235-1@student.aiub.edu

***Abstract*—This study explores the use of eight machine learning models—Logistic Regression, SVM, Random Forest, Extra Trees, KNN, XGBoost, LightGBM, and CatBoost—to detect Alzheimer's disease based on clinical and demographic data. The models were evaluated using accuracy, precision, recall, F1 score, AUC-ROC, and training time. Results indicated that tree-based ensemble models, particularly LightGBM and CatBoost, achieved the highest performance, with LightGBM showing the best balance between accuracy (95.58%), AUC-ROC (0.955), and training speed. XGBoost and Random Forest also demonstrated strong predictive capabilities, while KNN underperformed, with the lowest accuracy (71.86%) and AUC-ROC (0.75), indicating significant classification challenges. A paired t-test confirmed statistically significant differences between models, with LightGBM consistently outperforming others. Overall, LightGBM is recommended for Alzheimer's disease detection due to its superior performance and efficiency, while KNN is unsuitable for this task. These findings offer valuable insights into selecting effective machine-learning models for medical diagnosis, particularly in Alzheimer's detection.**

***Index Terms*—Alzheimer's disease, machine learning, Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbors, predictive models**

I. Introduction

Alzheimer's disease (AD) is one of the most common forms of dementia, affecting millions of people worldwide [1]. Its early diagnosis is vital for managing the disease, yet clinical diagnosis is often delayed due to subtle early-stage symptoms. With advancements in machine learning, there is an increasing interest in utilizing these technologies to predict and diagnose Alzheimer's disease from cognitive data [2].

Machine learning models offer promising capabilities to analyze large datasets and identify patterns associated with Alzheimer's disease [3]. Among the various approaches, supervised learning methods like Logistic Regression, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) have shown potential for accurate classification of Alzheimer's and non-Alzheimer's cases. However, the performance of these models can vary significantly depending on the dataset used, the features selected, and the algorithm's configuration.

The aims to compare the performance of Logistic Regression, Random Forest, SVM, and KNN in predicting Alzheimer's disease. Using a dataset comprised of cognitive features, the study evaluates the accuracy, precision, recall, and F1 score of each model. Additionally, the paper discusses the implications of these findings and offers insights into which machine learning model may be most appropriate for Alzheimer's detection.

II. Literature Review

Machine learning increasingly aids in early Alzheimer's detection. Models like Support Vector Machines (SVM) and Random Forests analyze brain scans and cognitive tests, with SVM known for accuracy and Random Forests for identifying complex patterns. However, limited data poses challenges, prompting researchers to explore deep learning for enhanced predictions.

J. Maroco et al. examine various statistical and machine learning methods for predicting the progression from Mild Cognitive Impairment (MCI) to dementia. Early dementia detection is crucial, so the authors compare seven classifiers—Multilayer Perceptron (MLP), Radial Basis Function Neural Networks (RBF), SVM, Classification and Regression Trees (CART), CHAID, QUEST, and Random Forests (RF)—against traditional methods like Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Logistic Regression (LR) to identify the most accurate forecasting methods for MCI progression [4].

This study highlights neuropsychological tests in diagnosing Alzheimer's Disease (AD) and how machine learning can improve these tests' accuracy. While important for detecting cognitive decline, neuropsychological assessments often fail to predict the transition from MCI to Alzheimer's. The research explores various machine learning models, including Neural Networks and decision tree-based approaches, to determine which are most effective in supporting early diagnosis [5].

Vijay S. Nori et al. focus on utilizing large-scale healthcare claims data to predict dementia onset using machine learning techniques. Despite the challenges of messy claims data, the

study aims to identify effective models that can manage data complexity and uncover key factors signaling dementia onset [6].

S. R. Bhagya Shree et al. emphasize the need for early Alzheimer's detection using machine learning on neuropsychological data. Early diagnosis is critical for managing disease progression and enhancing patient quality of life, as there is no cure for AD [7].

Karim Haddada et al. address the tension between the efficiency and explainability of AD diagnosis models, analyzing various machine learning techniques based on multiple features, including clinical and demographic data. Their findings reveal that clinical data is the most effective, while tree-based classifiers balance accuracy and usability. They underscore the importance of explainability in medical machine learning applications [8].

Vijay S. Nori's team also investigates machine learning applications for predicting Alzheimer's Disease and related dementias (ADRD) using administrative and electronic health records. Their "label learning" approach corrects inaccuracies in dementia case labels, enhancing predictive effectiveness for earlier diagnosis and interventions [9].

Eugene Y.H. Tang et al. discuss recent advancements in predicting dementia risk using machine learning, emphasizing the need for accurate risk assessment models amid rising global dementia cases. The paper reviews various approaches, noting their strengths and areas for improvement [10].

The review mentions studies employing cognitive tests like the Mini-Mental State Examination (MMSE) and Clinical Dementia Rating (CDR) for Alzheimer's classification. Traditional techniques struggle with cognitive decline complexities, while machine learning offers more dynamic, precise solutions. Some studies suggest cognitive data alone can be comparable to neuroimaging for early AD diagnosis, although performance drops in later stages [2].

Overall, machine learning, especially using SVM and Random Forests, effectively aids in early Alzheimer's diagnosis through brain scans (MRI) and cognitive tests (like MMSE). While some studies found SVM more accurate, others favored RF or ensemble models like XGBoost. The importance of careful data selection and model tuning to enhance performance is highlighted, along with challenges like small or imbalanced datasets, necessitating models that are accurate and easy for clinicians to trust.

III. Data Description

The Alzheimer's Disease Dataset provided by Rabie El Kharoua is a comprehensive repository of health-related data for 2,149 patients, each uniquely identified by an ID ranging from 4751 to 6900. It is designed for research and analysis of Alzheimer's Disease, encompassing a wide range of variables across several feilds [11].

- Demographic Details:
  Age: The age of the patients ranges from 60 to 90 years.
  Gender:

0	Male
1	Female

  Ethnicity:

0	Caucasian
1	African American
2	Asian
3	Other

  Education Level:

0	None
1	High School
2	Bachelor's
3	Higher

- Lifestyle Factors:
  BMI: Body Mass Index of the patients, ranging from 15 to 40. Smoking:

0	No
1	Yes

  Alcohol Consumption: Weekly alcohol consumption in units, ranging from 0 to 20.
  Physical Activity: Weekly physical activity in hours, ranging from 0 to 10.
  Diet Quality: Diet quality score, ranging from 0 to 10.
  Sleep Quality: Sleep quality score, ranging from 4 to 10.

- Medical History:
  For the following medical history features:
  Family History of Alzheimer's, Cardiovascular Disease, Diabetes, Depression, Head Injury, and Hypertension.

0	No
1	Yes

- Clinical Measurements:
  Systolic BP: Systolic blood pressure ranging from 90 to 180 mmHg.
  Diastolic BP: Diastolic blood pressure ranging from 60 to 120 mmHg.
  Cholesterol Total: Total cholesterol levels range from 150 to 300 mg/dL.
  Cholesterol LDL: Low-density lipoprotein cholesterol levels, ranging from 50 to 200 mg/dL.
  Cholesterol HDL: High-density lipoprotein cholesterol levels, ranging from 20 to 100 mg/dL.
  Cholesterol Triglycerides: Triglycerides levels range from 50 to 400 mg/dL.

- Cognitive and Functional Assessments:
  MMSE: Mental State Examination score, ranging from 0 to 30. Lower scores indicate cognitive impairment.
  Functional Assessment: Functional assessment score, ranging from 0 to 10. Lower scores indicate greater impairment.
  ADL: Activities of Daily Living score, ranging from 0 to 10. Lower scores indicate greater impairment.
  For Memory Complaints and Behavioral Problems,

0	No
1	Yes

- Symptoms:
  For Confusion, Disorientation, Personality Changes, Difficulty Completing Tasks and Forgetfulness,‘

0	No
1	Yes

- Diagnosis Information:
  For diagnosis,

0	No
1	Yes

- Confidential Information:
  Doctor In Charge: This column contains confidential information about the doctor in charge, with ”XXXConfid” as the value for all patients.

IV. Methodology

A. Data Collection and EDA

The data set utilized within this research came from [11] and contained detailed clinical and demographic information about patients. The dataset was inspected for missing values using the isnull() function from pandas. No significant proportion of missing data was found in any critical column, and hence, no major imputation was required. No outliers were found in the data either. Next, the ”DoctorInCharge” column as it was irrelevant to the explanatory analysis, and it intervened with running the machine models because it was the only column with an ”Object” datatype. In contrast, the rest of the data in the data set was of a numeric datatype. The ”PatientID” column was removed due to its irrelevance to our research. The remaining attributes were kept for subsequent research, while the ”Diagnosis” feature was identified as a classification target variable.

B. Feature Selection and Engineering

The dataset was partitioned into independent variables (features) and the dependent variable (target). Specifically, all columns except ”Diagnosis” were designated as features (denoted as X), while ”Diagnosis” served as the target variable (denoted as y). To ensure uniformity and enhance model performance, feature scaling was applied to selected models sensitive to feature magnitude. Standardization was performed using the StandardScaler, transforming the data into a mean of zero and a standard deviation of one.

C. Model Selection

We chose a mixture of common and uncommon machine-learning algorithms to assess if these algorithms would reliably and effectively predict Alzheimer's and Alzheimer's disease. The models were:

- Logistic regression.
- Support vector machine (SVM).
- Random Forest Classifier.
- Extra Trees Classifier.
- K-Nearest Neighbors (KNN).
- XGBoost.
- LightGBM.
- CatBoost.

These models were selected to provide a comprehensive comparison across linear models, ensemble methods, and gradient boosting frameworks.

D. Training and Evaluation

The dataset was divided into training and testing subsets using an 80-20 split, ensuring that 80% of the data was allocated for model training and 20% for evaluation. To maintain reproducibility, the split was randomized with a fixed seed (random state = 42).

A dedicated function, ‘train_evaluate_model,‘ was developed to streamline the training and evaluation process across different models. This function performs the following operations:

- Feature Scaling: This method applies standardization to the training and testing data for models sensitive to feature scaling (e.g., SVM, KNN, and Logistic Regression).
- Model Training: Fits the model on the training data.
- Prediction: Generates predictions and probability estimates on the testing data.
- Metric Calculation: Computes various performance metrics, including Accuracy, Precision, Recall, F1 Score, AUC-ROC, and Training Time.

The evaluation metrics were chosen to provide a holistic view of each model's performance, encompassing classification accuracy and the ability to discriminate between classes.

E. Statistical Analysis

To statistically assess the performance differences between models, 5-fold cross-validation was conducted, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was used as the primary evaluation metric. The cross-validation results were aggregated, and paired t-tests were performed to compare each model against the baseline Logistic Regression model. This analysis aimed to determine whether observed performance differences were statistically significant.

F. Tools and Libraries

All analyses were performed using Python 3.8, leveraging the following libraries:

- Pandas for data manipulation.
- NumPy for numerical operations.
- Scikit-learn for machine learning implementations and evaluation.
- Matplotlib and Seaborn for data visualization.
- LightGBM, CatBoost, and XGBoost for gradient boosting algorithms.
- SciPy for statistical testing.

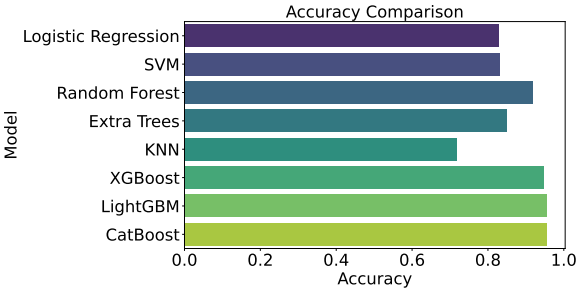
All libraries were utilized in standard configurations unless otherwise specified in the model selection subsection.

V. Results

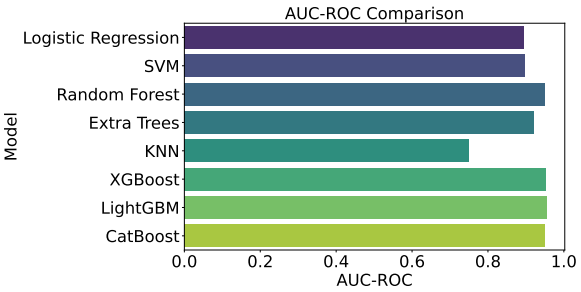
A. Model Performance Comparison

Eight models were compared based on six comparison matrices:

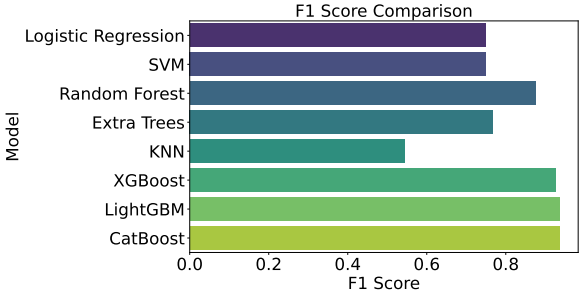
Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Training-Time
Logistic-Regression	0.8302	0.7899	0.7124	0.7491	0.8943	0.0131
SVM	0.8326	0.8000	0.7059	0.7500	0.8971	0.3988
Random-Forest	0.9349	0.9630	0.8497	0.9028	0.9496	0.3341
Extra-Trees	0.8488	0.8793	0.6667	0.7584	0.9314	0.2012
KNN	0.7186	0.6429	0.4706	0.5434	0.7491	0.0945
XG-Boost	0.9488	0.9580	0.8954	0.9257	0.9520	0.1053
Light-GBM	0.9558	0.9589	0.9150	0.9365	0.9547	0.0545
Cat-Boost	0.9558	0.9589	0.9150	0.9365	0.9511	1.0807



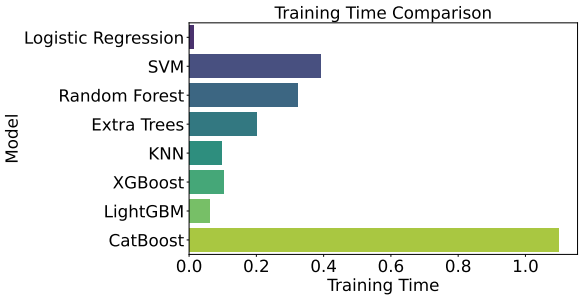
LightGBM and CatBoost had the highest accuracy score with XGBoost trailing slighly behind. KNN showed the lowest accuracy among all models.



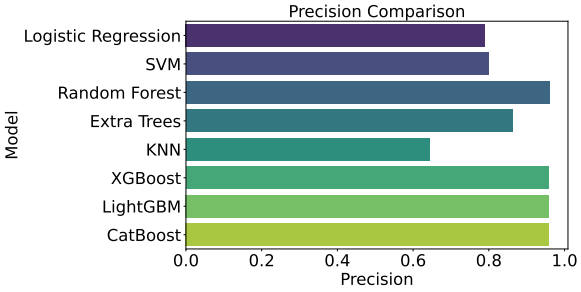
LightGBM had the highest AUC-ROC, which portrayed its ability to distinguish between positive and negative classes. XGBoost and Random Forest also showed excellent class separation with CatBoost slightly trailing behind.



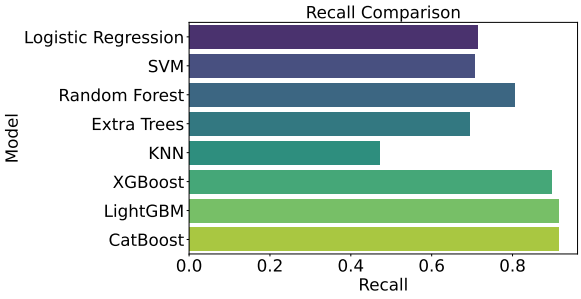
Light-GBM and CatBoost showed highest F1 score which meant a good balance between precision and recall. XGBoost followed by Random forest also shows good balance. KNN shows poor balance between precision and recall.



LightGBM was trained extremely fast which allowed for quick iterations. Logistic regression was also fast but not as efficient. CatBoost had the highest training time, which may limit its practicality in situations that require rapid model updates.



Random Forest had the highest precision which meant very few false positives. LightGBM and XGBoost followed Random Forest. KNN had the lowest precision indicating high false positive rates.



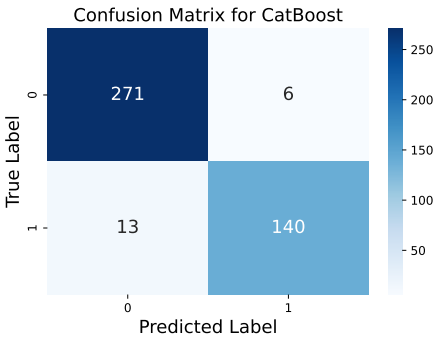
LightGMB and CatBoost both excelled in capturing true positives, followed by XGBoost and Random Forest. KNN struggled significantly, missing many true positives.

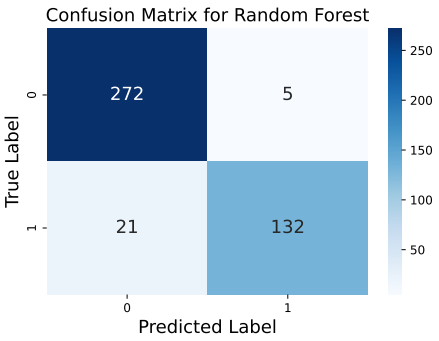
B. Confusion Matrices

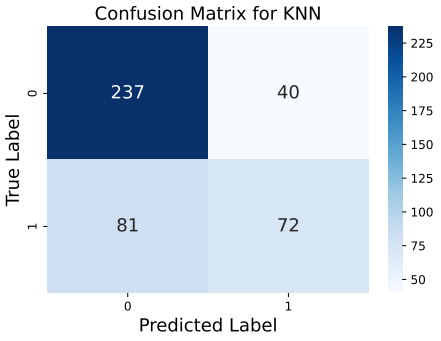
A confusion matrix breaks down a model's performance by comparing predicted outcomes to actual outcomes, providing counts of true positives, true negatives, false positives, and false negatives. It consists of four blocks:

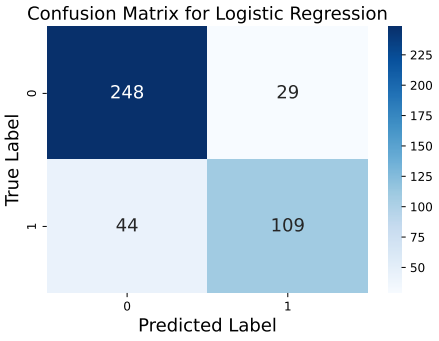
- **True Positives (TP)**: Correctly predicted positive cases.
- **True Negatives (TN)**: Correctly predicted negative cases.
- **False Positives (FP)**: Incorrectly predicted positive cases.
- **False Negatives (FN)**: Incorrectly predicted negative cases.

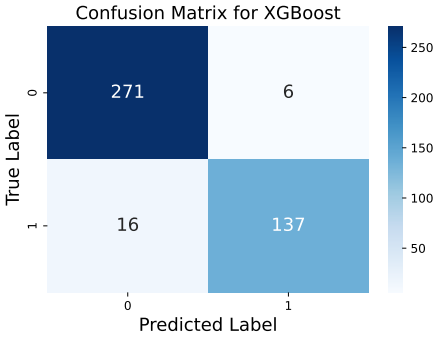
This matrix offers valuable insights into a model's identification accuracy.

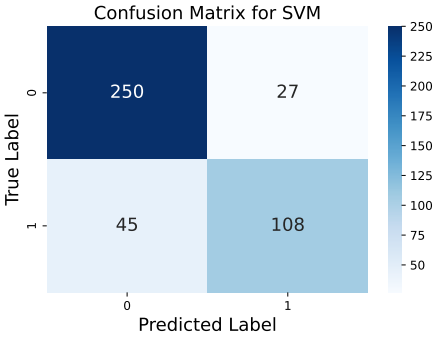


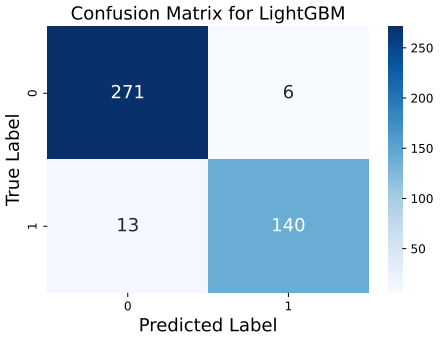


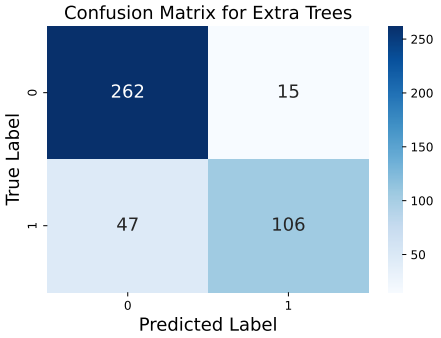












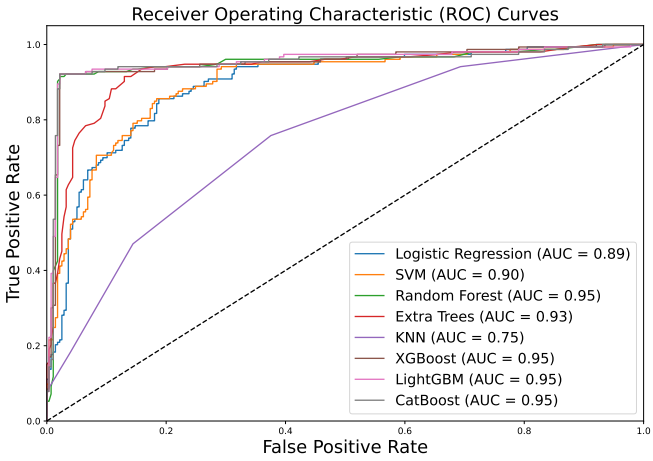
LightGBM and CatBoost showed the best performance, with high True Positives (TP) and True Negatives (TN) and very few False Positives (FP) and False Negatives (FN), making them highly accurate. XGBoost and Random Forest also performed well, with slightly more classification errors but still maintaining strong predictive power. The Extra Trees model was effective but had more misclassifications than the top models. Logistic Regression and SVM demonstrated moderate performance, struggling with some false classifications, while KNN was the weakest, showing the highest number of FP and FN, indicating it often misclassified cases.

Overall, ensemble models like LightGBM and CatBoost were the most reliable, while KNN was the least effective.

C. Receiver Operating Characteristic Curves

The ROC Curve Plot evaluates a model's classification performance by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at different thresholds, illustrating the model's ability to distinguish between positive and negative cases. The X-axis represents FPR, while the Y-axis represents TPR. A diagonal line indicates random guessing, and a curve closer to the top-left corner signifies

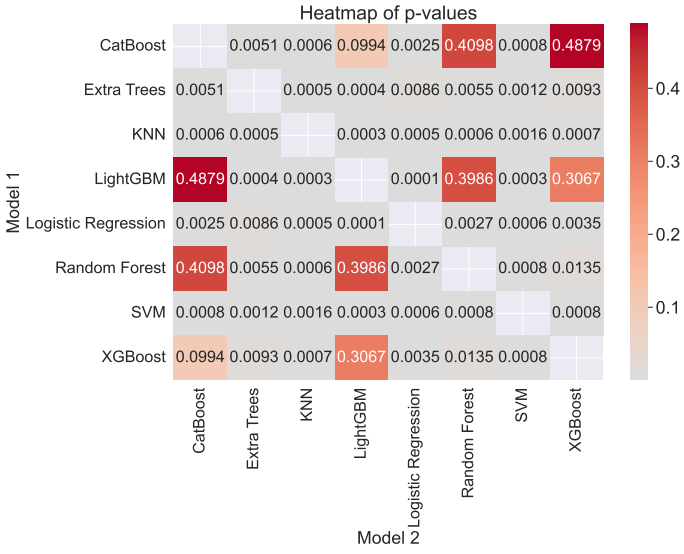
better accuracy. The Area Under the Curve (AUC) quantifies the model's overall effectiveness, with values near 1 indicating a highly effective model.

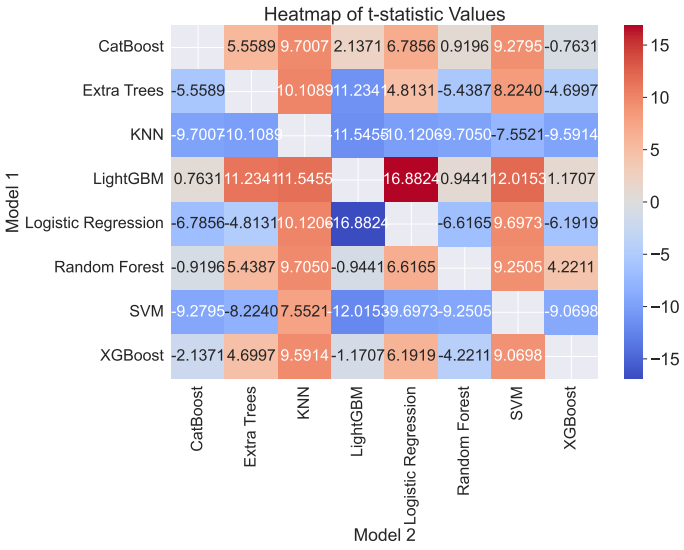


LightGBM, Random Forest, XGBoost, CateBoost have the highest AUC (Area Under the Curve= 0.95), indicating the most optimal classification ability. KNN (0.75) performed poorly, staying close to the diagonal, indicating almost random guessing.

D. Paired T-Test

The paired t-test is a statistical method employed to evaluate whether a significant difference exists between the means of two related groups. It is instrumental when the same subjects are measured under different conditions or times, such as before and after a treatment. The main objective of conducting a paired t-test is to assess if the mean differences between the two sets of observations are statistically significant, thus informing decisions based on empirical data. In comparing machine learning models, the paired t-test helps identify which models perform better relative to each other by analyzing the differences in their predictive accuracy on the same dataset, providing valuable insights into their relative effectiveness.





The comparison between Logistic Regression and LightGBM shows the highest t-statistic (16.88) and a very low p-value (7.22e-05), suggesting that LightGBM significantly outperforms Logistic Regression. Conversely, the comparison between Random Forest and LightGBM yields a p-value of 0.398, indicating no significant difference. Overall, LightGBM consistently performs against other models, while comparisons involving KNN also show notable differences, though not all are statistically significant.

E. Cross-Validated AUC Scores

Cross-Validated AUC Scores represent the Area Under the Curve (AUC) from Receiver Operating Characteristic (ROC) analysis. AUC measures a model's ability to distinguish between classes, with 1 being perfect classification and 0.5 being no better than random guessing. Cross-validation splits the data into several folds to evaluate the model's performance consistently across different subsets, ensuring more robust estimates of its generalization ability. The following Cross-Validated AUC scores were generated for this study:

Logistic-Regression	SVM	Random-Forest	Extra-Trees	KNN	XG-Boost	Light-GBM	Cat-Boost
0.9397	0.8405	0.9943	0.9638	0.5373	0.9920	0.9947	0.9989
0.9326	0.8338	0.9976	0.9646	0.5460	0.9963	0.9959	0.9952
0.9259	0.7777	0.9934	0.9708	0.5414	0.9917	0.9923	0.9949
0.9500	0.8584	0.9961	0.9652	0.5618	0.9954	0.9972	0.9969
0.7425	0.6533	0.7669	0.7589	0.5270	0.7638	0.7975	0.7676

Further, the scores were summarized: Tree-based models

Model	Mean AUC	Std AUC
Logistic Regression	0.8981	0.0874
SVM	0.7928	0.0836
Random Forest	0.9497	0.1022
Extra Trees	0.9247	0.0927
KNN	0.5427	0.0128
XGBoost	0.9479	0.1029
LightGBM	0.9555	0.0883
CatBoost	0.9507	0.1023

such as LightGBM, CatBoost, Random Forest, and XGBoost

had high AUC scores, with LightGBM achieving the highest mean AUC (0.956) and relatively low variability, indicating it as the best-performing model. CatBoost and Random Forest also performed well but exhibited higher standard deviations, suggesting that overfitting may have occurred. Logistic Regression and SVM were observed to perform moderately, while KNN underperformed with both low mean AUC and low standard deviation, indicating underfitting. Overall, LightGBM was identified as the top model due to its strong performance and consistency.

VI. Discussion

The comparison of eight machine learning models for detecting Alzheimer's disease reveals clear trends in performance across various evaluation metrics. Tree-based ensemble models—particularly LightGBM, CatBoost, Random Forest, and XGBoost—outperformed other models, showing higher accuracy, precision, recall, F1 scores, and AUC-ROC values. LightGBM consistently stood out as the best-performing model, with the highest accuracy (95.58%) and AUC-ROC (0.955), as well as a fast training time, making it both highly effective and efficient.

XGBoost and CatBoost also demonstrated strong performance but with higher training times compared to LightGBM, particularly CatBoost, which had the longest training time (1.08 seconds), potentially limiting its usability in time-sensitive applications. Random Forest also performed well, achieving high accuracy (93.49%) and excellent precision (96.3%), indicating a model well-suited for minimizing false positives.

On the other hand, KNN consistently underperformed in all metrics, with the lowest accuracy (71.86%), AUC-ROC (0.75), and F1 score (0.5434), indicating both underfitting and poor balance between precision and recall. Logistic Regression and SVM showed moderate results but lagged behind the ensemble models in most areas. While Logistic Regression trained quickly, it suffered from lower recall and F1 scores, indicating weaker classification performance.

The paired t-test analysis further solidified these findings, showing statistically significant differences between LightGBM and Logistic Regression. The comparison between Random Forest and LightGBM showed no significant difference, suggesting that these models may be similarly effective under certain conditions. KNN, in particular, exhibited significant performance differences compared to the stronger models, reinforcing its unsuitability for this task.

VII. Conclusion

In conclusion, the analysis highlights the superiority of tree-based ensemble models, particularly LightGBM, CatBoost, and XGBoost, for detecting Alzheimer's disease. LightGBM emerged as the top-performing model due to its high accuracy, precision, and recall and ability to distinguish between classes with minimal misclassification errors. Its quick training time also makes it a practical choice for applications requiring rapid model deployment and iteration. Conversely, models like KNN and Logistic Regression demonstrated significantly weaker performance, with KNN being especially ineffective due to its high misclassification rates and low AUC-ROC. As such, it is not recommended for this application. Overall, LightGBM is the best choice for Alzheimer's detection in this study, while Random Forest and XGBoost provide strong alternatives depending on specific use-case requirements. Future work could involve refining these models further or exploring other advanced techniques to enhance predictive accuracy and generalizability across broader datasets.

References

[1] John Hardy and Dennis J Selkoe. The amyloid hypothesis of alzheimer's disease: progress and problems on the road to therapeutics. *science*, 297(5580):353–356, 2002.

[2] Muhammad Irfan, Seyed Shahrestani, and Mahmoud Elkhodr. Early detection of alzheimer's disease using cognitive features: A voting-based ensemble machine learning approach. *IEEE Engineering Management Review*, 51(1):16–25, 2022.

[3] Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8:71792, 2014.

[4] Joao Maroco, Dina Silva, Ana Rodrigues, Manuela Guerreiro, Isabel Santana, and Alexandre de Mendonça. Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC research notes*, 4:1–14, 2011.

[5] HS Sheshadri, SR Bhagya Shree, and Murali Krishna. Diagnosis of alzheimer's disease employing neuropsychological and classification techniques. In *2015 5th International Conference on IT Convergence and Security (ICITCS)*, pages 1–6. IEEE, 2015.

[6] Vijay S Nori, Christopher A Hane, David C Martin, Alexander D Kravetz, and Darshak M Sanghavi. Identifying incident dementia by applying machine learning to a very large administrative claims dataset. *PLoS One*, 14(7):e0203246, 2019.

[7] SR Bhagya Shree and HS Sheshadri. An initial investigation in the diagnosis of alzheimer's disease using various classification techniques. In *2014 IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–5. IEEE, 2014.

[8] Karim Haddada, Mohamed Ibn Khedher, Olfa Jemai, Sarra Iben Khedher, and Mounim A El-Yacoubi. Assessing the interpretability of machine learning models in early detection of alzheimer's disease. In *2024 16th International Conference on Human System Interaction (HSI)*, pages 1–6. IEEE, 2024.

[9] Vijay S Nori, Christopher A Hane, William H Crown, Rhoda Au, William J Burke, Darshak M Sanghavi, and Paul Bleicher. Machine learning models to predict onset of dementia: a label learning approach. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 5:918–925, 2019.

[10] Eugene YH Tang, Stephanie L Harrison, Linda Errington, Mark F Gordon, Pieter Jelle Visser, Gerald Novak, Carole Dufouil, Carol Brayne, Louise Robinson, Lenore J Launer, et al. Current developments in dementia risk prediction modelling: an updated systematic review. *PloS one*, 10(9):e0136181, 2015.

[11] Rabie El Kharoua. Alzheimer's disease dataset, 2024.