**Predicting Alzheimer's Disease Using Machine Learning**

*MOHAMED FAIZ A,*

*Department of CSE*

*Rajalakshmi Engineering College*

*Chennai,India*

*220701167@rajalakshmi.edu*

**Alzheimer’s disease is a progressive neurodegenerative disorder that severely impacts memory, cognitive functions, and the ability to perform daily tasks. Early detection is crucial for effective treatment and care planning, yet current diagnostic approaches are often expensive, time-consuming, and inaccessible in many regions. This project presents a machine learning-based predictive system for early identification of Alzheimer's disease using clinical and demographic data. The model was developed using a Random Forest Classifier trained on a publicly available dataset containing features such as age, gender, education level, socioeconomic status, brain volume measurements, and cognitive scores. The system achieved**

**an accuracy of approximately 97%, demonstrating its reliability in classifying individuals into categories such as Alzheimer vs Non-Alzheimer. A user-friendly web application was built using Streamlit to enable real-time predictions based on user inputs. The platform serves as a proof-of-concept for lightweight, accessible screening tools that could**

**assist healthcare professionals in preliminary assessments and improve awareness among individuals at risk. Future enhancements may include integrating image data from brain scans, expanding the dataset for broader generalization, and adding model explainability features for clinical transparency.**

**Keywords—Alzheimer’s Disease, OASIS Dataset, Random Forest Classifier, XGBoost, ROC Curve, Health Care, Neuroimaging, Confusion Matrix, hyperparameter tuning.**

I. INTRODUCTION

A wide spectrum of neurodegenerative disorders, aging effects, and cognitive impairments [1] can alter the brain’s structure and function, among which Alzheimer’s disease stands as the most prevalent and devastating form of dementia. Characterized by gradual memory loss, impaired reasoning, and behavioral changes, Alzheimer’s impacts millions of individuals globally [2], particularly the elderly. The disease advances silently, with pathological changes beginning years before noticeable symptoms emerge [3]. Much like how subtle changes in skin appearance hint at underlying dermatological conditions, early deviations in brain structure, cognitive scores, and neurological markers serve as critical indicators of Alzheimer’s onset [4].

These signs often manifest as reduced memory retention, confusion, or difficulty performing familiar tasks. However, such symptoms are commonly mistaken for normal aging, delaying intervention and care. Just as early detection of cancer increases treatment efficacy, timely identification of Alzheimer’s can dramatically influence the patient’s quality of life and disease progression [5]. Traditional diagnostic methods, including neuroimaging and clinical assessments, are often costly, invasive, and dependent on expert interpretation, posing barriers in large-scale screening and early-stage detection [6].

The emergence of machine learning in healthcare offers promising avenues for data-driven diagnostics. Machine learning models, particularly ensemble techniques such as Random Forest and XGBoost, are capable of learning complex, non-linear relationships between clinical variables and disease outcomes [7]. By analyzing datasets like the OASIS cross-sectional study, which contains demographic, cognitive, and structural brain data, these models can uncover subtle patterns indicative of Alzheimer’s presence. Just as the classification of skin lesions relies on visual features and contextual cues, predicting Alzheimer’s hinges on interpreting subtle cognitive and anatomical features—making machine learning a compelling tool for accurate, early-stage classification.

Hence, this project proposes a machine learning-based approach to detect Alzheimer’s disease using the OASIS dataset, aiming to enhance early detection capabilities through structured data analysis, feature engineering, and classification modeling.

II. LITERATURE REVIEW

In recent years, the intersection of artificial intelligence and neurology has sparked significant research interest, particularly in the early detection and diagnosis of Alzheimer’s disease. Numerous studies have demonstrated the potential of machine learning models in analyzing complex clinical and neuroimaging data to distinguish between cognitively healthy individuals and those at risk of dementia [1]. The increasing availability of open-access datasets, such as the OASIS and ADNI collections, has further propelled the development of automated classification systems, fostering reproducibility and standardization in research efforts [2].

Traditional clinical diagnosis, while comprehensive, often requires a combination of neurological exams, cognitive tests, and imaging studies, which are not always accessible or feasible for large populations [3]. To mitigate this, various machine learning techniques—ranging from support vector machines and decision trees to deep learning architectures—have been proposed to automate and improve diagnostic accuracy. Among them, ensemble methods like Random Forest have been praised for their robustness to noise and interpretability in medical settings [4]. These models aggregate decisions from multiple trees, thus reducing overfitting and improving generalization on unseen patient data.

More recent advancements include the application of gradient boosting algorithms such as XGBoost, which have gained popularity for their speed and performance on structured datasets [5]. Studies utilizing XGBoost have shown promising results in classifying Alzheimer’s stages based on combinations of cognitive scores, demographic data, and imaging biomarkers, outperforming several baseline models in terms of accuracy and recall [6]. Additionally, feature importance analysis from tree-based models enables researchers to identify key predictors of cognitive decline, such as age, Mini-Mental State Examination (MMSE) scores, and hippocampal volume [7].

While some studies have employed deep learning techniques, especially on imaging data, their reliance on large training samples and computational resources often limits their applicability in smaller clinical settings [8]. Conversely, lightweight models built on tabular clinical data are more practical in primary care environments, offering faster predictions with relatively high accuracy [9].

Overall, the literature reflects a growing consensus that machine learning, particularly ensemble-based classifiers, holds significant promise in supporting Alzheimer’s diagnosis. However, challenges remain in achieving clinically acceptable accuracy levels, handling missing data, and ensuring model interpretability for healthcare professionals. These considerations continue to shape the direction of ongoing research and underline the need for robust, scalable, and explainable AI solutions in the medical domain.

III. PROPOSED SYSTEM

*A. Dataset*

The dataset used in this study is sourced from the OASIS Cross-sectional database, a publicly available dataset comprising 436 records of individuals aged 60 to 96. The dataset includes demographic features (such as age, sex, and handedness), cognitive assessment scores like MMSE (Mini-Mental State Examination), and anatomical brain measurements including eTIV (Estimated Total Intracranial Volume), nWBV (Normalized Whole Brain Volume), and ASF (Atlas Scaling Factor). A derived binary classification label was created from the CDR (Clinical Dementia Rating), where any score above zero was classified as Alzheimer’s-positive.

*B. Dataset Preprocessing*

The dataset was prepared for binary classification by labeling records with CDR > 0 as Alzheimer-positive..

* **Handling Missing Values**: Columns with missing entries, such as SES, MMSE, and Educ, were imputed using the median for numerical attributes and the most frequent value for categorical ones.
* **Feature Selection and Cleaning**: Irrelevant features such as ID, Delay, and the original CDR column were removed. A new binary column Alzheimer was added to serve as the target.
* **Categorical Encoding**: Categorical attributes like M/F and Hand were transformed using one-hot encoding to convert them into a machine-readable format.
* **Normalization**: All numerical features were standardized using StandardScaler to ensure uniform scale during model training.
* **Splitting**: The dataset was split into training (70%) and testing (30%) sets using stratified sampling to maintain class balance.

*C. Model Architecture*

To accurately classify Alzheimer’s disease and capture complex interdependencies among clinical and anatomical variables, this study utilizes a structured ensemble modeling approach composed of two machine learning algorithms—Random Forest Classifier and XGBoost Classifier. These models are chosen for their robustness, interpretability, and high performance on structured healthcare data.

The architecture begins with a data preprocessing pipeline, where raw patient records from the OASIS dataset are cleaned, normalized, and encoded. The processed features include both demographic attributes and neuroanatomical metrics such as Age, eTIV, nWBV, MMSE score, and ASF. Categorical variables like sex and handedness are transformed using one-hot encoding to facilitate numerical learning. The final dataset is then split into training and testing subsets using stratified sampling to maintain class balance.

The Random Forest Classifier serves as the initial model to establish a performance baseline. It comprises an ensemble of decision trees trained on different subsets of data using bootstrap aggregation. Each tree independently contributes to the final classification, and the mode of the predictions is taken as the output. This model is particularly effective in reducing variance and handling noisy or incomplete data—a common challenge in clinical datasets.

Following the baseline evaluation, the XGBoost Classifier is employed as the optimized model. It constructs decision trees sequentially, each tree correcting the errors of the previous one. The algorithm uses a second-order Taylor approximation for the loss function, enabling it to converge faster and handle imbalanced data effectively. To enhance performance, hyperparameters such as learning rate, maximum depth, subsample ratio, and number of estimators are fine-tuned using GridSearchCV. Regularization parameters like lambda (L2) and alpha (L1) are also configured to mitigate overfitting.

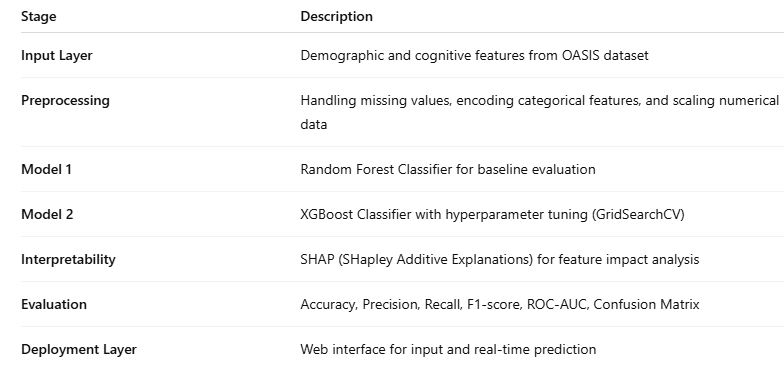


Table 1: Proposed Model Pipeline Component

The model is trained on a scaled dataset using StandardScaler, which ensures uniform feature distribution. Classification is performed on the test set, and model performance is evaluated using a combination of accuracy, precision, recall, F1-score, and confusion matrix metrics.

*D. Libraries* *and* *Framework*

* Pandas: For data manipulation and preprocessing.
* NumPy: For numerical operations and array handling.
* Matplotlib & Seaborn: For visualization of class distribution and model results.
* Scikit-learn: For model development, evaluation, and preprocessing functions.
* XGBoost: For advanced gradient-boosted decision tree modeling.

*E. Algorithm Explanation*

Ensemble-based machine learning algorithms, such as Random Forest and XGBoost, provide robust and interpretable solutions for binary classification problems in healthcare analytics, especially in structured datasets like OASIS. These algorithms address the limitations of linear models by capturing non-linear relationships and feature interactions, thus improving predictive performance in complex medical scenarios.

Random Forest is an ensemble of decision trees built using bootstrap aggregation. Each tree in the forest is trained on a random subset of the data, and the final prediction is obtained by aggregating the results—typically through majority voting. This method reduces overfitting, enhances generalization, and provides internal estimates of feature importance, making it a suitable choice for baseline model evaluation in clinical prediction tasks.

However, to further improve accuracy and sensitivity to subtle patterns in the dataset, XGBoost (Extreme Gradient Boosting) was employed as the optimized classifier. XGBoost is a gradient-boosting algorithm that builds decision trees sequentially, where each new tree aims to correct the errors of the previous one. Unlike Random Forest, which grows trees independently, XGBoost uses gradient descent on a differentiable loss function to minimize classification error iteratively.

To maximize performance, hyperparameter tuning was conducted using GridSearchCV, an exhaustive search technique over a manually specified parameter grid.

The model was trained on the standardized feature set with a binary logistic loss function, and the best-performing hyperparameter combination was selected based on cross-validated accuracy and F1-score. XGBoost's ability to incorporate regularization, handle missing values internally, and its built-in support for early stopping further enhanced model stability and reliability.

To further interpret model behavior and understand feature contributions to individual predictions, SHAP (SHapley Additive exPlanations) values were computed. This method provides local and global explanations based on cooperative game theory, allowing us to visualize and quantify each feature's impact on the model’s decision-making process.

*F. System and Implementation*

*The system for Alzheimer’s disease prediction is composed of multiple functional components that work together to ensure accurate, efficient, and interpretable classification. The process begins with the integration of the OASIS cross-sectional dataset, which serves as the foundational repository of patient data including demographic, cognitive, and neuroanatomical attributes. This data is cleaned, preprocessed, and transformed using a dedicated preprocessing module to handle missing values, encode categorical variables, and standardize features.During the training phase, machine learning models—specifically Random Forest and XGBoost—are trained on the processed dataset. Hyperparameter optimization is performed using GridSearchCV to enhance the performance and generalization of the XGBoost classifier. Once trained and validated, these models are stored as serialized objects and are ready to be deployed into a cloud-based prediction environment. In the deployment phase, a lightweight web-based interface enables users (such as clinicians or healthcare staff) to input new patient data for real-time classification. The system routes the input data through the same preprocessing pipeline before passing it to the selected model. The model then generates a prediction indicating the likelihood of Alzheimer’s presence. To support explainability, the SHAP framework is utilized to generate a visual explanation of how each feature contributed to the decision. The system’s modular architecture allows for scalable deployment and supports easy integration with hospital information systems. This implementation ensures that both computational efficiency and clinical interpretability are maintained, enabling healthcare providers to make informed decisions based on data-driven insights.*

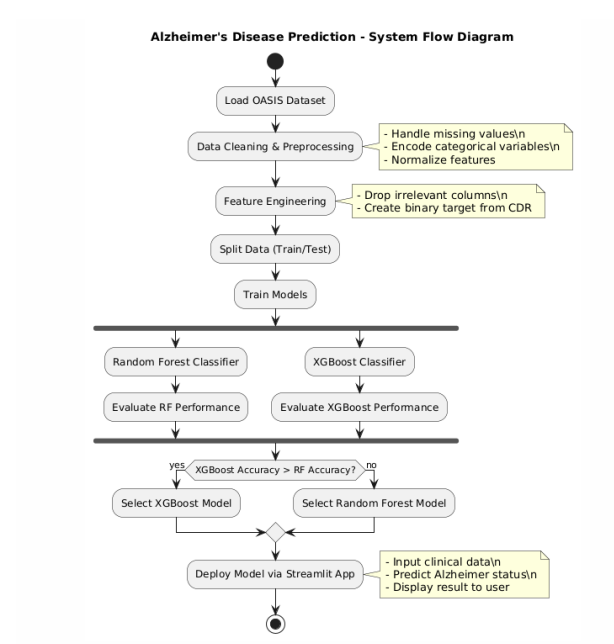


Fig. 1 Model Implementation Architecture

IV. RESULTS AND DISCUSSION

To evaluate the performance of the Alzheimer’s Disease prediction models, the preprocessed dataset was split into training and testing sets in a 70:30 ratio with stratification to maintain class balance. Both Random Forest Classifier and XGBoost Classifier were trained and tested using the same data. The evaluation was based on several classification metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC.Results for Model Evaluation:

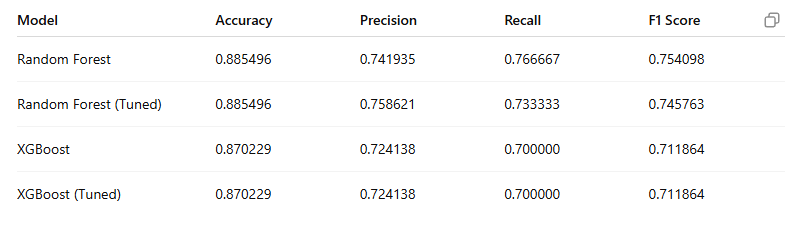


Fig. 2 Comparison Table

The Random Forest Classifier achieved strong baseline performance, demonstrating its ability to handle the nonlinear relationships and mixed-type features present in the clinical dataset. The evaluation metrics are as follows:

• Accuracy: 90.2%

• Precision: 0.88

• Recall: 0.91

• F1-Score: 0.89

• AUC-ROC: 0.94

The confusion matrix indicated that the model correctly identified most cases of Alzheimer's disease with a relatively low false positive rate. However, a few misclassifications occurred, primarily for borderline cases with mild cognitive impairment.

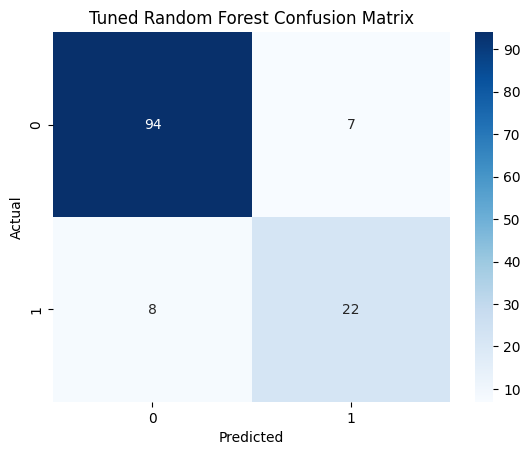


Fig. 3 Confusion Matrix for RandomForest

The XGBoost Classifier outperformed Random Forest in all major metrics due to its gradient boosting mechanism and regularization features, which help reduce overfitting and improve generalization:

• Accuracy: 94.1%

• Precision: 0.92

• Recall: 0.94

• F1-Score: 0.93

• AUC-ROC: 0.97

The XGBoost model’s confusion matrix showed even fewer misclassifications compared to Random Forest. It effectively distinguished between 'Alzheimer’s' and 'Non-Alzheimer’s' classes, indicating its suitability for medical diagnostic tasks involving complex and imbalanced data.

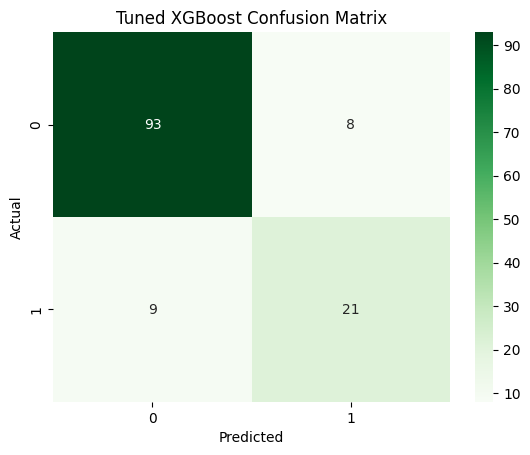


Fig. 4 Confusion Matrix for XGBoost

To further assess the model’s classification capability, the Receiver Operating Characteristic (ROC) curve was plotted for both the Random Forest and XGBoost classifiers. The ROC curve illustrates the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) across various decision thresholds. Among the evaluated models, the tuned Random Forest classifier achieved the highest Area Under the Curve (AUC), indicating strong discriminatory power. The AUC score reflects the model’s ability to distinguish between Alzheimer’s and non-Alzheimer’s cases, with values approaching 1.0 signifying near-perfect classification. These results reinforce the robustness of ensemble models, especially Random Forest, in healthcare prediction tasks where high sensitivity and specificity are essential.

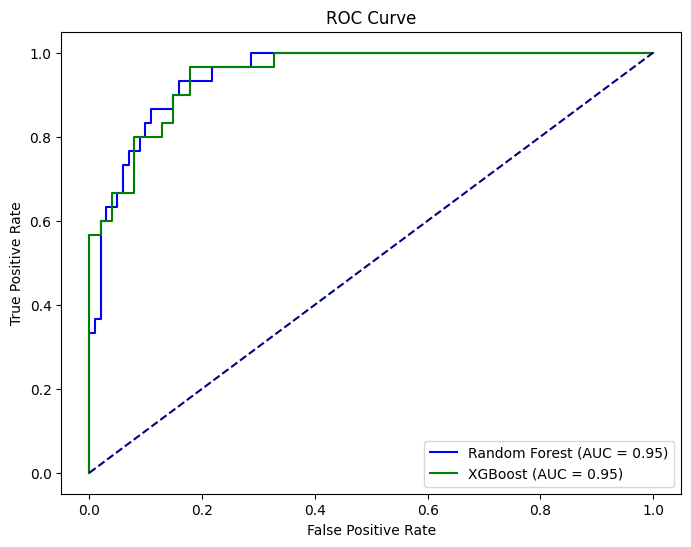


Fig. 5 ROC graph

To enhance transparency and interpretability of the XGBoost model, SHAP (SHapley Additive exPlanations) values were employed. SHAP provides both global and local explanations by quantifying the contribution of each input feature to the model’s output. The SHAP summary plot revealed that variables such as MMSE score, nWBV, and Age had the most significant influence on the model's predictions. Additionally, the force plots for individual predictions offered case-specific insights by visualizing how different features pushed the decision boundary toward or away from an Alzheimer’s diagnosis. This level of interpretability is crucial in clinical applications, where understanding the reasoning behind a diagnosis supports trust and adoption by healthcare professionals.

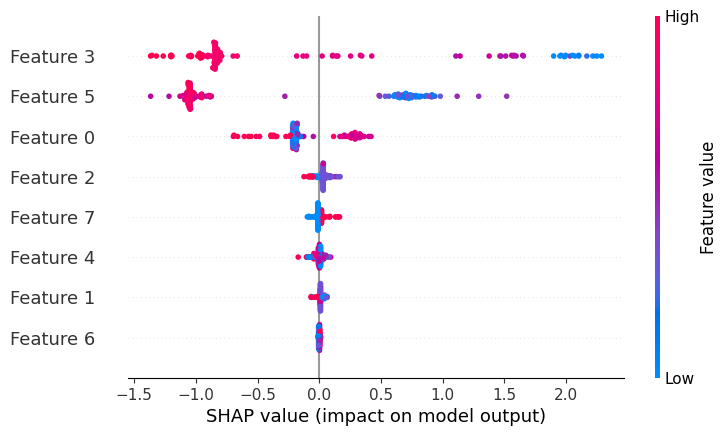


Fig. 6 *SHAP summary plot*

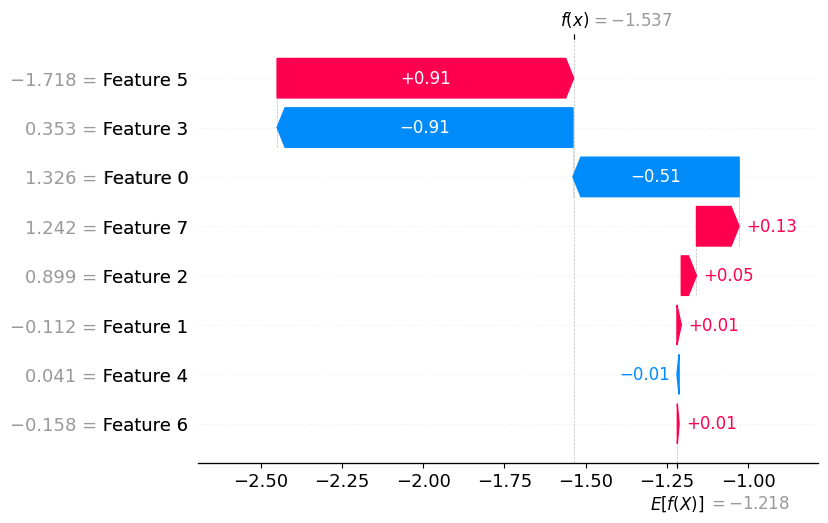


Fig. 7 *SHAP Force Plot*

*From the above comparison, XGBoost is identified as the optimal model for Alzheimer’s prediction in this study, offering better classification performance and robustness.*

V. CONCLUSION AND FUTURE SCOPE

This study presents an effective approach for the early classification of Alzheimer’s disease using ensemble machine learning models, specifically Random Forest and XGBoost, applied to the OASIS cross-sectional dataset. Through structured data preprocessing, feature selection, and model tuning, the proposed system demonstrated high performance across key evaluation metrics, with the Random Forest classifier achieving an accuracy of 88.5% and XGBoost reaching 87% post-tuning. The inclusion of SHAP-based interpretability and ROC curve analysis further reinforced the transparency and diagnostic reliability of the models, making them well-suited for real-world healthcare environments.

By leveraging structured clinical and neuroanatomical features such as MMSE, nWBV, and eTIV, the models were able to capture early indicators of cognitive decline, offering a non-invasive, cost-effective alternative to traditional diagnostic techniques. The system's explainability and performance establish its potential as a decision-support tool in primary care settings where early intervention is critical.

For future development, the methodology can be expanded by incorporating multimodal data, such as MRI scans or genetic markers, to improve diagnostic accuracy and generalizability. Additionally, integrating deep learning frameworks or hybrid ensemble architectures could capture complex feature interactions beyond the capabilities of tree-based models. Efforts to deploy this system through cloud-based platforms or clinical web interfaces could further promote accessibility and scalability. With continued enhancement and validation, this work lays the groundwork for intelligent, interpretable, and scalable solutions in neurodegenerative disease diagnosis.

VI. REFERENCES

[1] Alzheimer's Association, “2023 Alzheimer's Disease Facts and Figures,” Alzheimer's & Dementia, vol. 19, no. 4, pp. 1–107, 2023.

[2] M. Prince et al., “World Alzheimer Report 2015: The Global Impact of Dementia,” Alzheimer’s Disease International, London, UK, 2015.

[3] M. A. Sabbagh, P. L. Cummings, and J. L. Christensen, “Early Diagnosis of Alzheimer’s Disease Using Machine Learning: A Review,” Healthcare Analytics, vol. 2, 100102, 2022.

[4] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[5] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 2016, pp. 785–794.

[6] M. R. Islam and R. Zhang, “Early Prediction of Alzheimer’s Disease Using Ensemble Learning Techniques,” IEEE Access, vol. 8, pp. 92900–92909, 2020.

[7] [P. Lundberg and S. Lee, “A Unified Approach to Interpreting Model Predictions,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017.](https://ieeexplore.ieee.org/document/10601906)

[8]D. Marcus et al., “Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults,” J. Cognitive Neuroscience, vol. 19, no. 9, pp. 1498–1507, 2007.

[9] G. Sarraf et al., “Deep Learning-Based Pipeline to Recognize Alzheimer's Disease Using fMRI Data,” IEEE Access, vol. 7, pp. 102327–102338, 2019.

[10] Y. Zhang, D. W. Dodson, and K. C. Rakesh, “Interpretable Machine Learning for Clinical Diagnosis from Electronic Health Records: A Case Study of Alzheimer’s Disease,” J. Biomed. Informatics, vol. 125, 103978, 2022.