A Machine Learning Framework for Alzheimer’s Disease Detection: A Random Forest Approach with OASIS Data.

Neha Ishwar Reshmi [1], Swetha C S [2]

[1] Student, Department of MCA, Bangalore Institute of Technology, Karnataka, India (1BI23MC081)
[2]Assistant Professor, Department of MCA, Bangalore Institute of Technology, Karnataka, India

---***---

ABSTRACT

Alzheimer’s Disease (AD) is a progressive neurodegenerative disorder that causes memory loss, cognitive decline, and behavioral changes, making it a major global health challenge. With over 55 million people affected worldwide, timely and accurate diagnosis is crucial to improve patient outcomes. Traditional diagnostic methods such as MRI scans and clinical evaluations, while reliable, are often costly, time-consuming, and require expert involvement. Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) offer alternative approaches by detecting hidden patterns in patient data for early prediction. In this study, the OASIS Longitudinal Dataset, containing MRI, demographic, and cognitive features, was utilized to develop a predictive model for Alzheimer’s detection. A Random Forest Classifier was employed due to its robustness in handling heterogeneous data and its ability to provide feature importance insights. After preprocessing and training, the model achieved high accuracy in classifying subjects into non-demented, mildly demented, and moderately demented groups, with Clinical Dementia Rating (CDR), age, and MMSE scores identified as key predictors. The results demonstrate that Random Forest offers a reliable and interpretable solution for Alzheimer’s prediction, supporting its role as a clinical decision-support tool.

Keywords: Alzheimer’s Disease, OASIS Dataset, Random Forest Classifier, Machine Learning, Early Detection, Dementia Prediction.

I. INTRODUCTION

Alzheimer’s Disease (AD) is the leading cause of dementia worldwide, accounting for nearly 70% of dementia cases. It is a progressive neurological disorder that results in memory loss, impaired reasoning, and behavioral changes, ultimately leading to loss of independence. According to the World Health Organization (WHO), over 55 million people are currently living with dementia, and this figure is expected to rise significantly in the coming decades due to an aging global population.

Alzheimer’s Disease (AD) is the leading cause of dementia worldwide, accounting for nearly 70% of dementia cases. It is a progressive neurological disorder that results in memory loss, impaired reasoning, and behavioral changes, ultimately leading to loss of independence. According to the World Health Organization (WHO), over 55 million people are currently living with dementia, and this figure is expected to rise significantly in the coming decades due to an aging global population.

Conventional diagnostic techniques such as MRI scans, neuropsychological tests, and clinical assessments like the Mini-Mental State Examination (MMSE) and Clinical Dementia Rating (CDR) are widely used. However, these methods are expensive, time-consuming, and require trained professionals, limiting their use in large-scale screenings. In addition, traditional evaluations may introduce subjectivity, further delaying intervention.

Machine Learning (ML) offers a promising alternative by automatically analyzing patient data and detecting subtle patterns that are not easily visible to human experts. Ensemble methods like Random Forest are particularly effective in healthcare tasks, as they can manage high-dimensional data, avoid overfitting, and highlight important predictors. In this research, the OASIS Longitudinal Dataset is used with a Random Forest classifier to predict Alzheimer’s disease stages, providing a scalable and reliable framework that can support clinicians in early diagnosis and patient management.

II. LITERATURE SURVEY

Research on Alzheimer’s Disease (AD) prediction has gained significant attention in recent years, with machine learning and deep learning techniques being widely explored to enhance early diagnosis.

Marcus et al. [1] introduced the OASIS dataset as a benchmark resource for Alzheimer’s studies, enabling researchers to test various predictive models on real patient data. Breiman [2] demonstrated the effectiveness of Random Forest in handling complex medical datasets, providing robust classification results. Sarraf and Tofighi [3] highlighted how artificial intelligence could serve as a supportive diagnostic tool for clinicians by applying CNNs on fMRI data.

For advanced modeling, Dubey et al. [4] proposed machine learning methods for predicting the progression from mild cognitive impairment (MCI) to AD using MRI features. Zhang et al. [5] and Gupta et al. [6] applied ensemble and hybrid ML models to clinical datasets, showing that tree-based algorithms often outperform simple classifiers in dementia classification. Ali et al. [7] designed a framework for multimodal Alzheimer’s detection, proving that integrating imaging and demographic features improves predictive accuracy.

Comparative reviews have also guided the selection of algorithms. Sharma et al. [8] compared several classifiers and found Random Forest and SVM to be superior for Alzheimer’s screening tasks. Similarly, Yıldız et al. [9] emphasized that ML-based systems can serve as effective early screening tools for dementia, reducing delays in clinical evaluation. Beyond MRI and clinical scores, Liu et al. [10] explored speech-based analysis for AD detection, while Suk et al. [11] combined imaging with demographic features, offering a more comprehensive approach.

Recent works have also tested more complex modalities. Eslami et al. [12] explored deep learning–based networks using brain imaging, while Khan et al. [13] introduced hybrid ML-DL frameworks for improved classification. Singh et al. [14] and Kumar et al. [15] evaluated various ML methods, stressing that model accuracy depends heavily on proper preprocessing, balanced datasets, and feature selection.

Overall, the literature shows a clear trend: demographic and clinical data such as age, CDR, and MMSE are widely available and effective, but combining them with imaging or multimodal data leads to more accurate and robust Alzheimer's detection systems.

III. PROPOSED METHEDOLOGY

The proposed system focuses on predicting Alzheimer's Disease (AD) using clinical and imaging features available in the OASIS Longitudinal Dataset. By leveraging demographic information, cognitive scores, and MRI-based attributes, the model benefits from structured clinical data that reflect both biological and cognitive aspects of dementia progression. The complete methodology is divided into three main workflows: Dataset Collection, Data Preprocessing, and Model Development using the Random Forest Classifier.

3.1 Dataset Collection

Behavioral Dataset:

The dataset used in this study is the **OASIS Longitudinal Dataset**, which contains longitudinal MRI scans and associated clinical data of 150 subjects aged between 60 and 96 years. Each subject underwent multiple visits, providing valuable information for tracking disease progression. Along with MRI imaging features, the dataset includes **demographic attributes** (such as age, gender, and years of education) and **clinical attributes** (including Mini-Mental State Examination (MMSE) scores, Clinical Dementia Rating (CDR), and socioeconomic status). These features provide a reliable basis for machine learning–based Alzheimer's prediction. Importantly, the dataset is publicly available, anonymized, and widely used in Alzheimer's research, ensuring both quality and ethical compliance.

3.2 Data Preprocessing

The raw dataset could not be directly fed into machine learning models due to missing values, categorical variables, and class imbalance. Therefore, several preprocessing steps were performed:

1. Cleaning the Data – Irrelevant attributes such as subject IDs were removed since they do not contribute to prediction.

2. Handling Missing Values – Incomplete records were either imputed using the most frequent value or removed to maintain dataset integrity.

3. Encoding Categorical Variables – Features such as gender and education were encoded into numerical form using label encoding and one-hot encoding as appropriate.

4. Balancing the Dataset – Since the dataset contained more non-demented subjects compared to demented ones, balancing techniques such as SMOTE (Synthetic Minority Oversampling Technique) were applied to ensure fair training.

5. Splitting the Dataset – The dataset was divided into 80% training data and 20% testing data to evaluate model generalization.

After preprocessing, the dataset was ready to be used for classification with the Random Forest model.

3.3 Classification of Algorithms

Before finalizing the Random Forest Classifier (RFC), we compared several algorithms including Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). Each model was trained and evaluated on the processed OASIS dataset to measure its accuracy and reliability.

1.**Logistic Regression (LR):** Performs well on binary classification but struggles with the complex non-linear relationships present in clinical data One important part of machine learning is checking whether the model is overfitting or underfitting.

2. Naïve Bayes (NB): Simple and fast but less effective when attributes are correlated, as is common in Alzheimer's datasets.

3. Support Vector Machine (SVM): Strong in high-dimensional spaces, but parameter tuning is critical and interpretability is limited.

4. **K-Nearest Neighbors (KNN):** Easy to understand but highly sensitive to noise and class imbalance, reducing reliability.

5. Random Forest Classifier (RFC): An ensemble learning algorithm that constructs multiple decision trees and averages their predictions. It is robust, reduces overfitting, handles both categorical and numerical data, and provides feature importance for interpretability.

In our experiments, Random Forest achieved the highest accuracy and stability, making it the best choice for Alzheimer's prediction on the OASIS dataset.

3.4 Model Workflow

The workflow of the proposed Alzheimer's prediction system involves preprocessing the dataset, training the Random Forest Classifier, and evaluating performance using metrics such as accuracy, precision, recall, and F1-score. Feature importance analysis was also conducted to identify which clinical attributes (such as CDR, MMSE, and age) contributed most to the classification. This ensures that the model is both accurate and interpretable, making it a reliable decision-support tool for

clinicians.

IV. RESULT AND EVOLUTION

4.1 Model Result (Random Forest Classifier)

The Random Forest Classifier (RFC) applied to the OASIS Longitudinal Dataset achieved an accuracy of **96%**, with precision of **0.95**, recall of **0.96**, and an F1-score of **0.955**. These results are consistent with earlier works, which show that ensemble models such as Random Forest outperform simpler classifiers in Alzheimer's detection tasks [2, 5, 8]. The strong recall indicates that the majority of demented cases were correctly identified, which is crucial in healthcare applications, aligning with findings from Dubey et al. and Zhang et al. [4, 5]. Feature importance analysis highlighted that Clinical Dementia Rating (CDR)**,** Mini-Mental State Examination (MMSE)**,** and **age** were the most influential attributes, a trend also noted by Sarraf and Suk et al. [3, 11]. Preprocessing steps such as handling missing values, balancing the dataset with SMOTE, and encoding categorical variables improved stability, consistent with recommendations from Gupta et al. and Singh et al. [6, 14]. Overall, the Random Forest classifier proved to be the most reliable model for Alzheimer's prediction, supporting observations in prior studies [7, 15].

4.2 Comparitive Model Results

For comparison, other classifiers including Logistic Regression, Naïve Bayes, SVM, and KNN were also tested. Logistic Regression achieved moderate accuracy (88%) but struggled with the dataset's non-linear relationships. Naïve Bayes was fast but less effective due to correlated features. SVM performed well (92% accuracy) but required extensive parameter tuning and lacked interpretability. KNN showed unstable results and sensitivity to imbalanced data. In contrast, Random Forest consistently outperformed these methods, achieving higher accuracy and providing interpretable feature importance, which makes it a suitable choice for clinical applications in Alzheimer's detection.

4.3 Evaluation Matrix

The evaluation matrix summarizes the performance of the tested classifiers across multiple metrics:

- **Accuracy:** RFC achieved the highest accuracy (96%), outperforming SVM (92%), Logistic Regression (88%), and KNN (85%).
- **Precision:** RFC maintained precision of 0.95, minimizing false positives. This is important for avoiding unnecessary concern in non-demented patients.
- **Recall:** RFC achieved a recall of 0.96, showing strong sensitivity in detecting true dementia cases, which is critical in clinical applications.
- **F1-Score:** With an F1-score of 0.955, RFC demonstrated a balanced trade-off between precision and recall, surpassing other models.

These results confirm that Random Forest provides a stable and accurate solution for Alzheimer's detection using structured clinical and demographic data.

4.4 Discussion

The results demonstrate that Random Forest is highly effective in predicting Alzheimer's Disease progression using the OASIS dataset. Its strong recall ensures that most dementia cases are correctly identified, while high precision minimizes false alarms, aligning with prior studies emphasizing the strength of ensemble methods [2, 8, 14]. Compared to other classifiers, RFC offers a better balance between accuracy and interpretability, making it more suitable for practical clinical use.

However, some challenges remain. Variability in the dataset, including missing values and demographic imbalance, affected model performance slightly, a limitation also noted in previous works [5, 12]. Preprocessing steps such as balancing with SMOTE and careful encoding helped mitigate these issues, improving generalization. Importantly, feature importance analysis confirmed that CDR, MMSE, and age are consistent key predictors across multiple studies, reinforcing their clinical relevance. Overall, the results validate Random Forest as a reliable tool for Alzheimer's prediction and support the growing trend of integrating machine learning into healthcare decision support systems.

V. FUTURE WORK

Future research can focus on expanding the dataset by including a larger and more diverse population. The current OASIS dataset is valuable but limited in terms of sample size and demographic variation. A larger dataset would improve the model's generalizability and reduce potential biases related to age, gender, and ethnicity.

Another important direction is the integration of multimodal data sources, such as MRI scans, genetic markers, and cognitive test results. Combining different modalities would allow the model to capture both structural brain changes and behavioral patterns, leading to more accurate and reliable predictions.

Deep learning models, particularly Convolutional Neural Networks (CNNs) and hybrid ML-DL frameworks, could be applied in future studies. These methods can automatically learn complex representations from raw MRI data, providing insights that traditional models may overlook.

Real-world deployment of Alzheimer's prediction systems in clinical decision support tools is another key area for future work. Building user-friendly platforms that can assist doctors during routine checkups would ensure that machine learning methods transition from research to practical healthcare use.

Finally, incorporating explainable AI (XAI) techniques would enhance transparency in model predictions. By highlighting which features contribute most to classification, XAI can improve clinician trust, making the system more interpretable and ethically aligned with healthcare standards.

VI. CONCLUSION

This research presented a machine learning–based approach for Alzheimer's Disease prediction using the OASIS Longitudinal Dataset and Random Forest Classifier. The model achieved strong performance, with an accuracy of 96%, demonstrating that ensemble methods are highly effective in analyzing clinical and demographic data for dementia classification.

The feature importance analysis revealed that attributes such as Clinical Dementia Rating (CDR)**,** Mini-Mental State Examination (MMSE)**,** and age played the most significant role in predicting Alzheimer's progression. These findings are consistent with prior studies, confirming the clinical relevance of these indicators and validating the model's reliability.

Overall, the study highlights the potential of machine learning as a supportive tool in healthcare. By offering faster, scalable, and cost-effective predictions, the proposed system can complement traditional diagnostic methods and enable early interventions. With further improvements through multimodal data integration and real-world validation, this approach can contribute significantly to advancing Alzheimer's care and research.

VII. REFERNCES

[1] Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). *Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle-aged, nondemented, and demented older adults*. Journal of Cognitive Neuroscience, 19(9), 1498–1507.

[2] Breiman, L. (2001). *Random forests*. Machine Learning, 45(1), 5–32.

[3] Sarraf, S., & Tofighi, G. (2016). *Classification of Alzheimer's disease using fMRI data and deep learning convolutional neural networks*. arXiv preprint arXiv:1603.08631.

[4] Dubey, S., Chakrabarti, P., & Singh, S. (2020). *Predicting progression from mild cognitive impairment to Alzheimer's disease using machine learning models*. Procedia Computer Science, 167, 1704–1711.

[5] Zhang, D., Wang, Y., Zhou, L., Yuan, H., & Shen, D. (2011). *Multimodal classification of Alzheimer's disease and mild cognitive impairment*. NeuroImage, 55(3), 856–867.

[6] Gupta, Y., Lama, R. K., & Kwon, G. R. (2019). *Prediction and classification of Alzheimer's disease based on combined features from apolipoprotein-E genotype, cerebrospinal fluid, MR, and FDG-PET imaging biomarkers*. Frontiers in Computational Neuroscience, 13, 72.

[7] Suk, H. I., Lee, S. W., & Shen, D. (2014). *Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis*. NeuroImage, 101, 569–582.

[8] Sharma, A., & Verbeke, W. (2020). *Ensemble learning for Alzheimer's disease classification from structural brain MRI data*. Applied Soft Computing, 100, 106960.

[9] Khan, A., Hussain, T., & Mehmood, I. (2020). *A hybrid deep learning model for Alzheimer's disease classification using MRI data*. Cognitive Neurodynamics, 14(6), 785–797.

[10] Singh, A., Kumar, A., & Singh, S. (2021). *Machine learning models for early diagnosis of Alzheimer's disease: A comparative study*. Biomedical Signal Processing and Control, 70, 102957.