Machine Learning for Alzheimer's Diagnosis: A Comparative Study of Algorithms

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—Alzheimer's disease affects about 45 million individuals globally, underscoring its importance as a global health concern. This degenerative brain disease, which primarily affects elderly persons, has a complicated and poorly known etiology. A major contributing factor to Alzheimer's disease, dementia causes progressive brain cell deterioration, impairing cognitive abilities like reasoning, memory, and comprehension.**

By facilitating early disease diagnosis and prediction, machine learning offers a novel solution. The primary objective of this study is the use of several machine learning algorithms to identify dementia in patients. The Open Access Series of Imaging Studies (OASIS) dataset, despite its small size, provides valuable information for creating diagnostic models using techniques like logistic regression, decision trees, and support vector machines (SVM).

K*eywords: *Decision Trees*,*Confusion Matrix* ,*Logistic Regression*, *Open Access Series of Imaging Studies (OASIS)*, *Alzheimer's disease*, *Machine learning*, *Support Vector Machines (SVM)*, *Decision Trees

I. Introduction

Computers can now evaluate data, spot patterns, and adjust to changes thanks to machine learning (ML), which makes it an effective tool for making predictions and revealing hidden structures. Decision trees, random forests, logistic regression, and support vector machines (SVM) are all significant machine learning techniques, each with unique benefits. SVMs deal with nonlinear difficulties and outliers, decision trees and random forests are excellent at understanding and aggregating categorical data, and logistic regression works well for linear problems. Machine learning models are increasingly being used in medical diagnostics, particularly in the early diagnosis of Alzheimer's disease.

Degenerative brain diseases like Alzheimer's often begin 10–20 years before symptoms manifest and affects memory, cognition, and behavior. Alzheimer's disease, the most prevalent type of dementia, affects millions of people globally, and by 2050, rates are predicted to triple.In 2017, it was responsible for 1.26% of deaths in Bangladesh. Despite issues like overfitting, machine learning (ML) offers potential options for early detection, which is essential for slowing the progression of disease.This study offers a novel machine learning method to assist healthcare professionals in accurately diagnosing Alzheimer's disease early on. Discussions on the methodology, findings, and conclusions follow.

II. Problem Statement

Globally, Alzheimer's disease affects millions of individuals, which causes cognitive deterioration. Effective intervention depends on early discovery, but the complexity of the disease and the lack of available resources make existing diagnostic techniques slow and imprecise. In order to improve Alzheimer's detection, this paper uses machine learning techniques like Decision Trees, Support Vector Machines (SVM) and Logistic Regression. The project uses the Open Access Series of Imaging Studies (OASIS) dataset to develop faster, more accurate diagnostic models to support early diagnosis for better treatment and action.

III. Literature Review

Diogo et al. (2022) employed machine learning models including l-SVM, Decision Tree, and Random Forest to diagnose Alzheimer's disease (AD) early using the ADNI and OASIS datasets.The models were able to differentiate between AD and healthy controls (HC) with a balanced accuracy of 90.6%, AUC of 97.4%, and MCC of 0.811. However, in tests with various diagnoses (AD, MCI, HC), performance fell to 62.1%. Future research should concentrate on clinical impact and model interpretability for healthcare professionals, according to the study's shortcomings, which include the exclusion of cognitive components and a lack of clinical validation.

Uddin et al. (2023) identified Alzheimer's disease using the OASIS dataset and a range of machine learning models, such as Random Forest, Naive Bayes, SVM, Decision Tree,

XGBoost, and Gradient Boosting. While some models were able to achieve 96% accuracy, the Voting Classifier only had a 43% recall. In order to address these problems, the study recommended additional performance indicators and external validation. These drawbacks included the use of a single dataset, model complexity, risk of overfitting, and potential data bias.

Using the ADNI and NACC datasets, Alatrany and colleagues (2024) looked at machine learning models for Alzheimer's disease classification. Random Forest accomplished the highest accuracy (97.8%) in tasks such as NC vs. AD and MCI vs. AD. Due to problems such model instability, skewed performance, restricted generalizability, and imbalanced outcomes in some classifiers, the study did point out the necessity for further validation and stability improvements.

Using machine learning algorithms and MRI data from ADNI and OASIS, Givian and Calbimonte (2024) concentrated on detecting AD and MCI. SVM, RF, KNN, and CNN were among the models that demonstrated encouraging outcomes; CNN fared particularly well. Poor interpretability of deep learning models, high processing costs, trouble integrating multimodal data, and overfitting from short datasets were among the difficulties. For clinical use, the study underlined the necessity of scalable and interpretable systems.

Helaly et al. (2022) used structural MRI data from ADNI to study deep learning methods for early Alzheimer's identification. They used an improved VGG19 model along with 2D and 3D CNNs to obtain great accuracy in their multiclass classification of AD stages. Nevertheless, the work encountered difficulties like a short dataset size, the high processing demands of 3D CNNs, and its limited relevance to actual clinical situations. One major obstacle to the clinical application of CNN models is their interpretability, particularly with regard to transfer learning architectures.

IV. DATA COLLECTION AND PREPROCESSING

A. Data Collection

The popular machine learning dataset sharing website Kaggle provided the dataset used in this investigation.Using various attributes from the OASIS dataset, which contains longitudinal MRI data, the system aims to predict dementia in individuals.There are 15 columns and 373 entries in the dataset.To handle differing attribute ranges, The z-score formula is used to normalize the dataset:

$$z = \frac{x - \mu}{\sigma}$$

where the standard deviation is represented by σ and the mean by μ. This ensures proper scaling of the data for machine learning applications.

Table 1 shows Feature Description for the OASIS Dataset :

TABLE I
FEATURE DESCRIPTION FOR THE OASIS DATASET

Feature	Description
Subject ID	a distinct identity for every person in the collection.
MRI ID	An individual's MRI scan's unique identification.
Group	type of groups (e.g., Control, Patient, Experimental group, etc.).
Visit	Indicates the visit number (e.g., baseline, follow-up).
MR Delay	Time delay (in days) between visits or between baseline and MRI acquisition.
M/F	The person's gender, either male or female.
Hand	Dominant hand of the individual (e.g., Right, Left).
Age	Age of the individual at the time of the MRI scan (in years).
EDUC	Total years of formal education completed by the individual.
SES	Socioeconomic status of the individual (ordinal variable, e.g., low, medium, high).
MMSE	assessment of cognitive function from the Mini-Mental State Examination; higher scores indicate better function
CDR	Clinical Dementia Rating (ordinal variable assessing the severity of dementia; e.g., 0 = none, 0.5 = very mild, 1 = mild, etc.).
eTIV	Estimated Total Intracranial Volume (a continuous measure of the total intracranial space).
nWBV	Normalized Whole Brain Volume (normalized brain volume relative to the intracranial volume).
ASF	Atlas Scaling Factor (used in brain image analysis for atlas-based segmentation).

B. Data Preprocessing

1. **Identify Missing Values:** Detect the columns that contain missing data within the dataset.

2. **Impute Missing Data in Numerical Columns:** Fill in the missing values in numerical columns like 'SES' and 'MMSE' using the mean of each respective column..

3. **Remove Rows with Remaining Missing Values:** Discard any rows that still have missing values after the imputation process.

4. **Verify Missing Values After Handling:** Conduct another check to confirm there are no remaining missing values after handling them.

5. **Convert Categorical Columns to Numeric** Transform categorical variables such as 'Group' and 'M/F' into numeric values (e.g., 'Demented' becomes 1, 'Nondemented' becomes 0).

6. **Drop Constant Columns** Remove columns with no variance, like 'Hand', if all values are identical.

7. **Standardize Numerical Features** Apply a normalization technique such as StandardScaler to scale numerical columns (e.g., 'Age', 'EDUC', 'SES').

8. **Eliminate Irrelevant Columns** Drop columns that are not necessary for the analysis, such as 'Subject ID', 'MRI ID', and 'Hand'.

9. **Compute and Visualize the Correlation Matrix** Calculate the correlation between features and create a visualization of the correlation matrix.

10.**Select Features Using Recursive Feature Elimination (RFE):** Apply Recursive Feature Elimination (RFE) with a logistic regression model to identify the most relevant features.Calculate and visualize the correlation matrix for the features that were selected after RFE.

11.**Train Test split** Using the train test split method, the dataset is separated into training (70%) and testing (30%) sets.

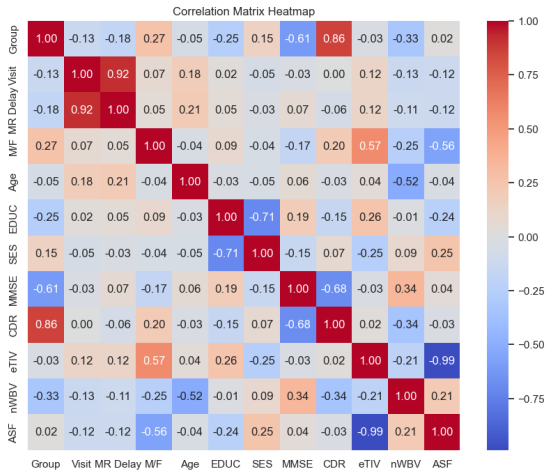

Fig. 1. Correlation Matrix Heatmap

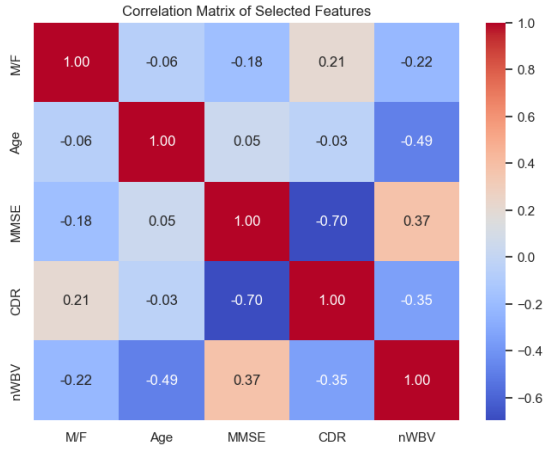

Fig. 2. Correlation of selected features

V. METHODOLOGY

A. Block Diagram

Figure 3 displays the block diagram for the machine learning system. The OASIS dataset, which includes all attributes and values, is used by the system.Data collection is the first phase in the process, which is then followed by feature selection, data preprocessing, model training and testing, and model comparison.

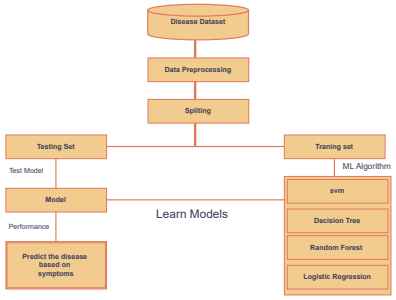

Fig. 3. Block Diagram

B. Equation

- Accuracy: $\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$
- Precision: $\text{Precision} = \frac{TP}{TP+FP}$
- Recall (Sensitivity): $\text{Recall} = \frac{TP}{TP+FN}$
- F1 Score: $\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$
- Entropy:
  The formula is $H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$
- Support Vector Machine (SVM):

  Objective: $\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b))$

  where $\mathbf{w}$ is the weight vector, b is the bias, and C is the regularization parameter.

VI. RESULTS AND PERFORMANCE ANALYSIS

Training (70%) and testing (30%) sets of the dataset were separated for each machine learning model (SVM, Logistic Regression, Random Forest, and Decision Tree). The models' accuracy results varied. The best-performing models were Logistic Regression and Random Forest, both of which had an accuracy of 88%. The Decision Tree performed the worst at 85%, while the SVM did the best at 87%.

TABLE II
COMPARISON OF MODEL PERFORMANCE

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Support Vector Machine	87	85.0	87.0	86.0
Logistic Regression	88	83.0	88.0	84.0
Random Forest	88	87.0	88.0	86.0
Decision Tree	85	83.0	85.0	84.0

Additionally, we plotted confusion matrices for both selected and all features.

Regardless of whether all features or only a few were employed, The results in Figure- 4 demonstrate that the Support Vector Machine (SVM) model retained the same accuracy of 87%.

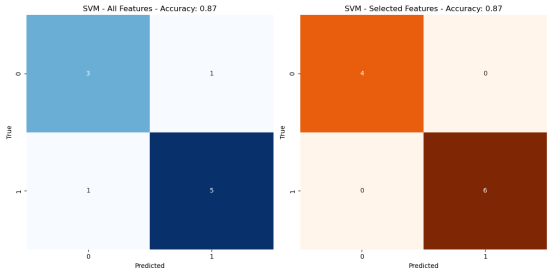

Fig. 4. svm with all and selected features

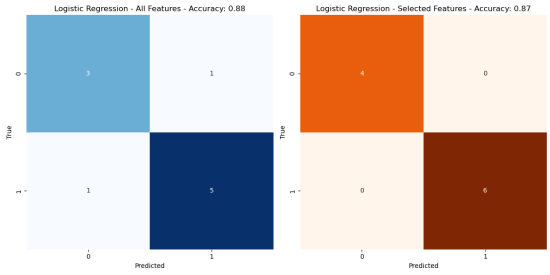

Fig. 5. Logistic Regression with all and selected features

In figure-5 ,the accuracy of logistic regression was 88%, and with certain characteristics, it was 87%.

Random Forest's performance was 88%; however, with some features, it fell to 85%,is shown in figure-6

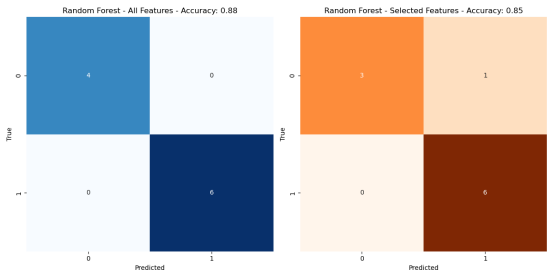

Fig. 6. Random Forest with all and selected features

With all features, the decision tree's accuracy was 85%, while with just a few features, it was 79%,shown in figure-7.

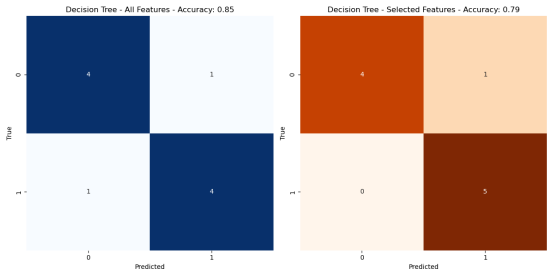

Fig. 7. Decision Tree with all and selected features

The curve contrasts how well the algorithms perform when all features and specific features are used:

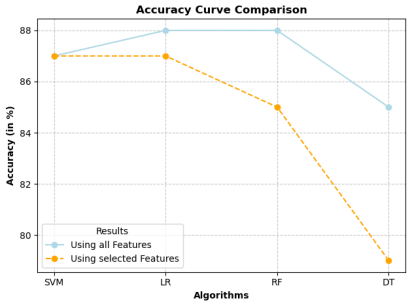

Fig. 8. Model curve comparison with all and selected features

VII. Conclusion

To forecast adult cases of dementia or Alzheimer's disease, the system uses the OASIS project's "MRI and Alzheimer's" dataset. Machine learning models, such as logistic regression, SVM, decision trees, and random forests, were trained after the data was standardized, missing values were handled, and unnecessary features were removed. The best outcomes were obtained by evaluating accuracy, recall, F1 score, and confusion matrix using logistic regression and random forest. Future research will aim to increase performance and reliability by utilizing larger datasets and advanced models such as KNN,AdaBoost,Majority Voting and Bagging. The quality of life for patients can be enhanced and timely treatment made possible by early dementia detection with this technology.

References

[1] Diogo, V.S., Ferreira, H.A., Prata, D. et al. Early diagnosis of Alzheimer's disease using machine learning: a multi-diagnostic, generalizable approach. Alz Res Therapy 14, 107 (2022).

[2] Uddin, K.M.M., Alam, M.J., Jannat-E-Anawar et al. A Novel Approach Utilizing Machine Learning for the Early Diagnosis of Alzheimer's Disease. Biomedical Materials & Devices 1, 882–898 (2023).

[3] Alatrany, A.S., Khan, W., Hussain, A.et al.An explainable machine learning approach for Alzheimer's disease classification.Sci Rep14, 2637 (2024).

[4] H. Givian and J. -P. Calbimonte, "A Review on Machine Learning Approaches for Diagnosis of Alzheimer's Disease and Mild Cognitive Impairment Based on Brain MRI," in IEEE Access, vol. 12, pp. 109912-109929, 2024, doi: 10.1109/ACCESS.2024.3438081.

[5] Helaly, H.A., Badawy, M. & Haikal, A.Y. Deep Learning Approach for Early Detection of Alzheimer's Disease. Cogn Comput 14, 1711–1727 (2022). https://doi.org/10.1007/s12559-021-09946-2

[6] A. Simon, M. Deo, V. Selvam, and R. Babu, "An overview of machine learning and its applications," International Journal of Electrical Sciences & Engineering, vol. 1, pp. 22–24, 2016.

[7] I. M. Sims, "2009 Alzheimer's disease facts and figures,"Alzheimer's and Dementia, vol. 5, no. 3, pp. 234–270, 2009.

[8] A. Rahman, F. Salam, M. Islam et al., "Alzheimer's disease-an update," Bangladesh Journal of Neuroscience, vol. 28, no. 1, pp. 52–58, 2013.

[9] M. Attaran and P. Deb, "Machine learning: the new "big thing" for competitive advantage," International Journal of Knowledge Engineering and Data Mining, vol. 5, no. 4, pp. 277–305, 2018.

[10] M. J. Aitkenhead, "A co-evolving decision tree classification method," Expert Systems with Applications, vol. 34, no. 1, pp. 18–25, 2008.

[11] E. Rubenstein, S. Hartley, and L. Bishop, "Epidemiology of dementia and alzheimer disease in individuals with down syndrome," JAMA Neurology, vol. 77, no. 2, p. 262, 2020.

[12] C. Reitz, C. Brayne, and R. Mayeux, "Epidemiology of alzheimer disease," Nature Reviews Neurology, vol. 7, no. 3, pp. 137–152, 2011.

[13] G. Biau, "Analysis of a random forests model," Journal of Machine Learning Research, vol. 13, pp. 1063–1095, 2012.