Undergraduate Conference on Intelligent Computing and Systems (UCICS 2025)
26-27 February, 2025; Varendra University, Rajshahi, Bangladesh

Investigating Different Machine Learning Techniques for Alzheimer's Disease Classification

Ankur Kumar Mondol
Dept. of Computer Science and Engineering
Pabna University of Science and Technology
Pabna-6600, Bangladesh
Email:ankurmondol2109@gmail.com

Shahriar Siddique Arjon
Dept. of Computer Science and Engineering
Pabna University of Science and Technology
Pabna-6600, Bangladesh
Email:aorjon123@gmail.com

Pabitra Kumar Biswas
Dept. of Computer Science and Engineering
Pabna University of Science and Technology
Pabna-6600, Bangladesh
Email:pk605532@gmail.com

Tamanna Yasmin
Dept. of Computer Science and Engineering
Pabna University of Science and Technology
Pabna-6600, Bangladesh
Email:tamannayasmin2632001@gmail.com

Nakib Aman
Dept. of Computer Science and Engineering
Pabna University of Science and Technology
Pabna-6600, Bangladesh
Email:nakib.cse@pust.ac.bd

S. M. Hasan Sazzad Iqbal
Dept. of Computer Science and Engineering
Pabna University of Science and Technology
Pabna-6600, Bangladesh
Email:sazzad@pust.ac.bd

***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 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, random forest, decision tree, and support vector machine (SVM).The best outcomes 88.00%, were obtained using random forest and logistic regression.**

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

I. Introduction

The study of algorithms and statistical models that let computers recognize patterns and draw conclusions without explicit programming is known as machine learning (ML). Machine learning enables the analysis of healthcare data to enhance disease detection and treatment [1]. Machine learning algorithms develop effectively with practice and adjust to changing conditions. A machine learning system that uses data to find patterns and forecast future events is called a model.For two-group classification, a support vector machine (SVM) is a dependable supervised learning model that works well with little data. A method for classifying data and modeling the connections between independent variables and a binary dependent variable is called logistic regression.A supervised classification technique called a decision tree divides data at internal nodes and makes predictions about target classes at leaf nodes.A flexible, user-friendly supervised learning technique, random forest frequently produces excellent results without the need for hyperparameter modification. Alzheimer's disease, a degenerative and irreversible brain disorder that impairs memory, thinking, and everyday functioning and frequently begins 10–20 years before symptoms manifest.It is being diagnosed using machine learning models.Dementia is the primary cause of Alzheimer's disease, impacting 40–50 million individuals worldwide, with an estimated 131.5 million by 2050.There is currently no cure for dementia, which is a deterioration in brain function that primarily affects older adults and affects behavior, emotions, and daily living.Bangladesh ranked 152nd in the world in 2017 with 9,917 deaths from Alzheimer's and dementia, which accounted for 1.26% of all deaths [2].Bangladesh has a low awareness of Alzheimer's disease, few resources, and an increasing risk because of obesity.Prompt identification and treatment may enhance the results of Alzheimer's disease.To help medical technicians diagnose Alzheimer's, a machine learning model has been developed that can accurately identify the disease in its early stages.The introduction is the first section of this article. The problem statement and literature review of Alzheimer's disease are the second and third sections, respectively. Dataset description and preprocessing technique, results, performance, and conclusion are covered in Sections 4 to 7.

II. Problem Statement

Millions of people worldwide suffer from Alzheimer's disease, which causes cognitive loss. Effective intervention depends on early discovery, yet because of the disease's complexity and the lack of available resources, current diagnostic techniques are frequently delayed and inaccurate. In order to improve Alzheimer's detection, this study investigates machine learning methods such as Decision Tree, Random Forest, Support Vector Machine (SVM), and Logistic Regression. This work intends to create quicker and more precise diagnostic models using the Open Access Series of Imaging Studies (OASIS) dataset, enabling early diagnosis for better care and treatment.

III. Literature Review

Alzheimer's disease (AD) is a degenerative illness characterized by behavioral abnormalities, memory loss, and cognitive impairment.This section includes review of the literature. Kumar and Azad [3] published a study of Alzheimer's disease (AD) research employing machine learning (ML) and deep learning (DL), evaluating several classifiers with feature selection to enhance accuracy. CNN models attained accuracy exceeding 90%, and SVM and ensemble approaches surpassed 85%. Metaheuristic methods such as genetic algorithms enhanced speed optimisation. Nonetheless, issues such as data imbalance and model interpretability remain. The bibliometric analysis underscores the increasing global research initiatives. The research emphasises the necessity of interdisciplinary collaboration and enhanced data integration to facilitate improved Alzheimer's disease diagnosis. In [4], Uddin et al.used a range of machine learning models and achieved 96% accuracy in Alzheimer's disease prediction; however, they faced problems with data bias, overfitting, and low recall (43%) in the Voting Classifier. They recommended better performance metrics and outside validation. In [5], Alatrany et al.used the ADNI and NACC datasets to classify Alzheimer's; Random Forest obtained 97.8% accuracy. Problems such model instability, distorted performance, and the need for more validation and stability were highlighted by them, nevertheless. In [6], Givian used MRI data and machine learning models such as SVM, RF, KNN, and CNN to detect AD and MCI; CNN yielded the best results. Problems like overfitting, poor interpretability, and high processing costs were highlighted in the study to highlight the need for scalable and interpretable systems in clinical application.After reviewing 116 studies, Malik et al. [7] examined deep learning methods for predicting Alzheimer's disease (AD). The accuracy of CNNs and ensemble learning was high; some models even surpassed 95%. The diagnosis was enhanced by multimodal data fusion and MRI feature extraction. There was widespread use of SVMs and deep learning models such as VGG16 and ResNet. Data imbalance, interpretability, and the requirement for sizable labeled datasets are among the difficulties. The study focuses on clinical adoption of explainable AI. Multimodal integration and enhancing model generalization should be the main goals of future studies.

IV. Dataset Description and Preprocessing

A. Dataset Description

The popular machine learning dataset sharing website Kaggle [8] 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.For a somewhat skewed and reasonably normal distribution of these missing data, the mean has been utilized.

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 normal-

ization 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.

Correlation Matrix and Correlation Matrix of selected features is shown in Fig.1 and Fig.2.

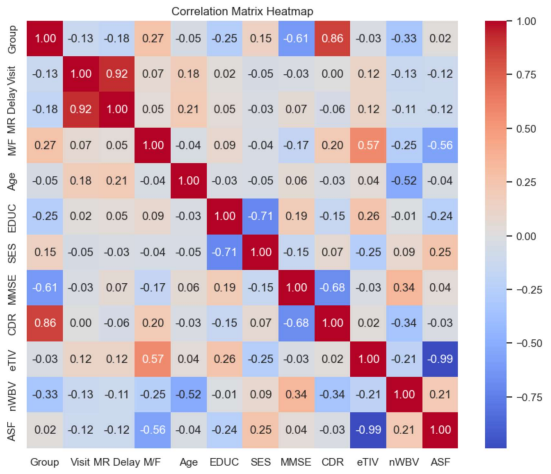

Fig. 1. Correlation Matrix

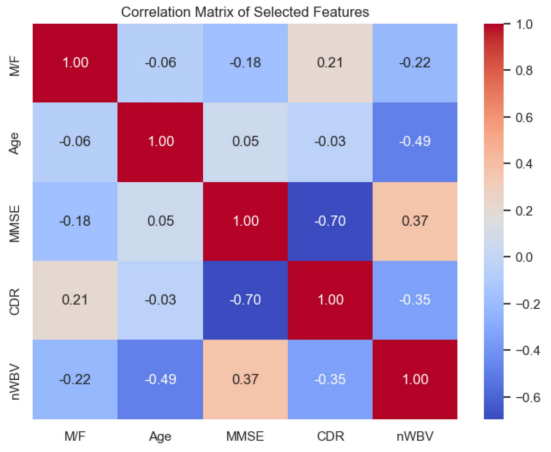

Fig. 2. Correlation of selected features

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

V. METHODOLOGY

In the machine learning system depicted in Fig. 3, preprocessing (including handling missing values, encoding, and feature selection using RFE) comes after data collection. Accuracy, precision, recall, and F1-score are used to assess performance when the dataset is divided into training and testing sets. The best model is then used to the prediction of Alzheimer's disease.

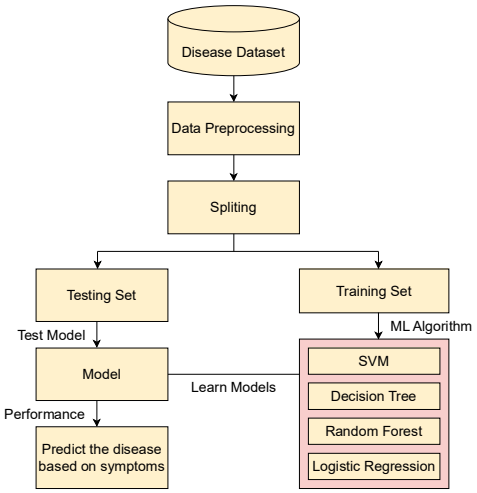

Fig. 3. Block Diagram

The four models used to classify Alzheimer's disease have various methods since they are based on various concepts.

A. Support Vector Machine(SVM)

SVM is a supervised learning method that applies one-vs-one or one-vs-rest for multi-class classification and separates classes using an ideal hyperplane and kernels while balancing errors with a regularization parameter. The equation for hyperplane is:

$$w \cdot x + b = 0 \tag{1}$$

The decision function for classification is:

$$f(x) = \text{sign}(w \cdot x + b) \tag{2}$$

B. Logistic Regression

A classification algorithm called Logistic Regression uses the sigmoid function to forecast probabilities. It solves multi-class jobs using one-vs-rest or softmax techniques, minimizes log-loss, and employs regularization to avoid overfitting. Applications such as risk analysis and medical(Alzheimer's disease) diagnosis make extensive use of it.Logistic regression predicts the likelihood of a binary result by employing the sigmoid function:

$$p(x) = \frac{1}{1 + e^{-(w \cdot x + b)}} \tag{3}$$

The probability output is always within the range $[0, 1]$.

C. Random Forest

Random Forest is an ensemble technique that uses random feature selection and bootstrapping to decrease overfitting and increase accuracy. It works well for regression and classification, including the diagnosis of Alzheimer's disease.

Formula for Classification (majority voting) and Regression (average of predictions) are:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_N(x)\} \tag{4}$$

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} T_i(x) \tag{5}$$

D. Decision Tree

A decision tree is a supervised learning model that uses factors like Gini Impurity, Entropy, and Mean Squared Error for classification and regression. Though it is sensitive to changes in the data and prone to overfitting, it can handle both continuous and categorical data. Random Forest and other ensemble approaches are based on it. The formula for Gini Impurity (Classification),Entropy (Classification),Mean Squared Error (MSE) (Regression)are:

$$G = 1 - \sum_{i=1}^{C} p_i^2 \tag{6}$$

Where p_i is the proportion of samples of class i in the node.

$$H = -\sum_{i=1}^{C} p_i \log_2(p_i) \tag{7}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{8}$$

Where y_i is the target value and $\bar{y}$ is the mean target value.

.

VI. RESULTS AND PERFORMANCE ANALYSIS

Training (70%) and testing (30%) sets of the dataset were separated for the SVM, Random Forest, Logistic Regression, and Decision Tree models.The best accuracy (88.00%)was attained by Random Forest and Logistic Regression, followed by SVM (87.00%),while the lowest accuracy (85.00%) was attained by Decision Tree.Table-2 shows Performance Comparison of Different Applied Machine Learning Models.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT APPLIED ML MODELS

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

In Fig.4,The curve contrasts how well the algorithms perform when all features and specific features are used.

The results in Fig.4,demonstrate that the Support Vector Machine (SVM) model retained the same accuracy of 87.00%.The accuracy of Logistic Regression was 88.00%, and with certain characteristics, it was 87.00%.Random Forest's performance was 88.00%; however, with some features, it fell to 85.00%.With all features, the Decision Tree's accuracy was 85.00%, while with just a few features, it was 79.00%.

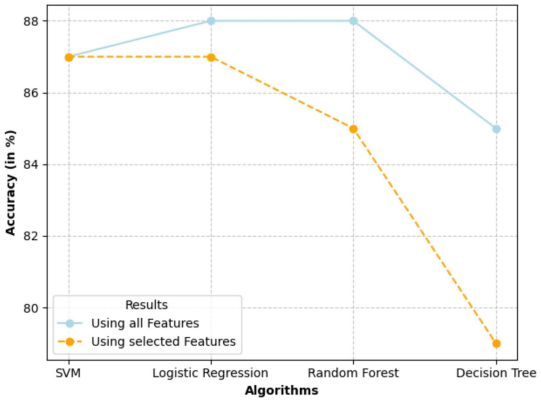

Fig. 4. Model accuracy curve comparison with all and selected features

VII. CONCLUSION

Following the preprocessing of the OASIS dataset, the system predicts dementia and Alzheimer's cases using machine learning models such as SVM, decision trees, logistic regression, and random forests. The best outcomes were obtained by evaluating accuracy, recall, F1 score, and confusion matrix using logistic regression and random forest.The limitations of the research include the use of limited techniques, tiny dataset, limitations of methodologies and a time-consuming process.The generalizability of the results may be affected by overfitting, bias, high variability, limited diversity, and limited feature representation that can occur in small datasets like OASIS.Future studies will use larger datasets and more advanced models like KNN,AdaBoost, Majority Voting,and Bagging in an effort to improve performance and reliability.Early dementia detection using this technology can improve patients' quality of life and enable prompt treatment.

REFERENCES

[1] H. Alshamlan, A. Alwassel, A. Banafa, and L. Alsaleem, "Improving alzheimer's disease prediction with different machine learning approaches and feature selection techniques," *Diagnostics*, vol. 14, no. 19, p. 2237, 2024.

[2] M. Bari Antor, A. S. Jamil, M. Mamtaz, M. Monirujjaman Khan, S. Aljahdali, M. Kaur, P. Singh, and M. Masud, "A comparative analysis of machine learning algorithms to predict alzheimer's disease," *Journal of Healthcare Engineering*, vol. 2021, no. 1, p. 9917919, 2021.

[3] R. Kumar and C. Azad, "Comprehensive overview of alzheimer's disease utilizing machine learning approaches," *Multimedia Tools and Applications*, pp. 1–53, 2024.

[4] K. M. M. Uddin, M. J. Alam, M. A. Uddin, and S. Aryal, "a novel approach utilizing machine learning for the early diagnosis of alzheimer's disease," *Biomedical Materials & Devices*, vol. 1, no. 2, pp. 882–898, 2023.

[5] A. S. Alatrany, W. Khan, A. Hussain, H. Kolivand, and D. Al-Jumeily, "An explainable machine learning approach for alzheimer's disease classification," *Scientific Reports*, vol. 14, no. 1, p. 2637, 2024.

[6] 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," *IEEE Access*, 2024.

[7] I. Malik, A. Iqbal, Y. H. Gu, and M. A. Al-antari, "Deep learning for alzheimer's disease prediction: A comprehensive review," *Diagnostics*, vol. 14, no. 12, p. 1281, 2024.

[8] J. Boysen, "Mri and alzheimer's dataset." https://www.kaggle.com/datasets/jboysen/mri-and-alzheimershttps://www.kaggle.com/datasets/jboysen/mri-and-alzheimers, 2024. Accessed: 2025-02-09.