Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

A Hybrid Approach for Harnessing the Power of Machine and Deep Learning for Precise Alzheimer's Diagnosis

Amna Shakeel
Department of Computer Science, COMSATS University Islamabad,
Wah Campus, Wah Cantt 47040, Punjab, Pakistan
Email: amnashakeel002@gmail.com

Rabbia Shakeel
Department of Computer Science, COMSATS University Islamabad,
Wah Campus, Wah Cantt 47040, Punjab, Pakistan
Email: rabbiashakeel13@gmail.com

Muhammad Ammar Khan
Department of Computer Science, HITEC University Taxila, 47080,
Punjab, Pakistan. Email: shkhammar2003@gmail.com

ABSTRACT

Dementia is a debilitating condition characterized by a progressive decline in memory and cognitive functions, primarily affecting individuals aged 60 and above. Alzheimer's Disease (AD), the most common degenerative form of dementia, specifically impairs memory and critical mental abilities due to the degeneration of brain cells and neural connections [1]. Currently, there is no cure for AD, and life expectancy post-diagnosis ranges from four to eight years, though some may survive up to two decades depending on various factors. According to the World Health Organization (WHO), approximately 50 million people worldwide suffer from dementia, with nearly 10 million new AD cases reported annually.

Keywords: Machine learning, Deep Learning, Alzheimer's Diagnosis

1. Introduction

Dementia is a debilitating condition characterized by a progressive decline in memory and cognitive functions, primarily affecting individuals aged 60 and above. Alzheimer's Disease (AD), the most common degenerative form of dementia, specifically impairs memory and critical mental abilities due to the degeneration of brain cells and neural connections [1]. Currently, there is no cure for AD, and life expectancy post-diagnosis ranges from four to eight years, though some may survive up to two decades depending on various factors. According to the World Health Organization (WHO), approximately 50 million people worldwide suffer from dementia, with nearly 10 million new AD cases reported annually. The risk of developing AD increases significantly with age, reaching 50% for those aged 85 and older [2]. Ultimately, AD damages brain regions responsible for vital functions such as breathing and heart rate, leading to death.

AD was first identified in 1906 by German neurologist and psychiatrist Alois Alzheimer, who studied a 51-year-old patient, Auguste D. Her family brought her to Dr. Alzheimer in 1901 due

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

to severe memory loss, speech difficulties, and behavioral changes. Over five years, she exhibited worsening symptoms, including confusion, aggression, and language impairment. After she died in 1906, an autopsy revealed significant brain shrinkage, fatty deposits in blood vessels, and the presence of neurofibrillary tangles and amyloid plaques, hallmarks of AD [3]. The condition was formally named in 1910 and entered medical literature in 1907. AD progresses through three stages: early, middle, and late, with symptoms worsening over time. Early signs include frequent forgetfulness, difficulty recalling names, poor decision-making, and trouble completing routine tasks [8]. While the exact cause remains unknown, factors such as genetics, aging, head injuries, hypertension, and lifestyle choices contribute to its development [5].

Neuroimaging techniques like MRI, CT, and PET scans play a crucial role in AD detection, with MRI being the most relied upon despite its high cost [4]. Machine learning (ML) and deep learning (DL) have shown promise in analyzing medical data for early diagnosis, but challenges such as dataset variability and algorithm interpretability limit clinical application [6]. Key research gaps include the need for reliable early biomarkers, a deeper understanding of metabolic and genetic factors, and improved differentiation between AD and other dementias. Combining multi-modal data (genetic, proteomic, metabolomic, and neuroimaging) could enhance diagnostic accuracy, though integration remains complex. Addressing these gaps is essential for advancing AD detection and treatment strategies.

1.1 Problem Statement

Alzheimer's is a neurodegenerative disease that begins with a moderate decline in cognitive functioning, worsens over time, and eventually leads to the death of brain cells [7]. Despite the progress made in medical science, accurately predicting AD remains challenging. Literature studies use MRI, PET, and CT scan images that are computationally expensive to handle, lack some features like demographic information and clinical tests of patients, and are only restricted to older patients. This research aims to enhance prediction reliability by using a less computationally intensive method, utilizing two numeric datasets that include more features, such as demographic information and clinical tests, and cover a broader age range, from 33 to 96-year-old patients. Using Deep Learning and Machine Learning techniques for extracting the best features and building trustworthy algorithms that can accurately diagnose AD, this research lends support to human diagnosis and conventional measurements in healthcare, and also assists physicians in making treatment decisions. Evaluation of prediction models will be performed through metrics like accuracy, precision, recall, F1-score, and AUC/ROC.

1.2 Aims and Objectives of Research

The main objectives of the study of Alzheimer's detection are as follows:

- To implement feature ranking techniques to identify the most important clinical variables that cause the development of AD using diverse datasets.
- To create prediction models of both ML and DL to detect Alzheimer's occurrence and progression accurately.
- To develop a diverse dataset that encompasses all the key features for detecting AD.
- To optimize performance metrics like accuracy and AUC/ROC, experiment with feature selection, hyperparameter adjustment, and ML and DL techniques.

By developing precise prediction models and decision support tools, the research aims to enhance the well-being of individuals affected by AD. The ultimate goal of this research is to contribute to advancements in healthcare.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

2. Literature Review

Alzheimer's disease (AD) is a progressive neurodegenerative disorder leading to cognitive decline, memory loss, and impaired daily functioning. Early and accurate detection remains challenging, necessitating advanced diagnostic approaches. Current methods include cognitive assessments (e.g., MMSE, MoCA), neuroimaging (MRI, PET, CT), and biomarker analysis (CSF, blood tests) [9]. Cognitive tests evaluate memory and problem-solving skills, while structural imaging (MRI, CT) detects brain atrophy. Functional imaging (PET, fMRI) assesses metabolic changes, and biomarker tests identify pathological proteins like amyloid-beta and tau. Despite their utility, these techniques vary in accuracy, cost, and invasiveness, highlighting the need for improved early detection strategies.

Positron Emission Tomography (PET) scans utilize radiotracers to measure glucose metabolism, revealing reduced activity in AD-affected brain regions (parietal, temporal cortices) even before symptom onset [10]. PET offers high diagnostic accuracy (~90%) but involves radioactive exposure and high costs [11]. Computed Tomography (CT) is more affordable and non-invasive but lacks sensitivity in early AD detection, primarily identifying advanced-stage atrophy and plaques [12]. Magnetic Resonance Imaging (MRI) excels in structural analysis, detecting hippocampal atrophy a key early AD biomarker with superior spatial resolution [13]. Advanced MRI variants, like sodium MRI, show promise in tracking neuronal death but require further validation [14]. While MRI is non-invasive and highly accurate, its expense and incompatibility with metallic implants limit accessibility.

Recent studies leverage Machine Learning (ML) and Deep Learning (DL) to enhance AD classification using neuroimaging data. Inertia tensor analysis combined with ML achieved 90% accuracy in staging AD from MRI scans [15]. Hybrid models like GAN-CNN (94% accuracy) and Multi-Task Deep Learning (97.1% accuracy) improve segmentation and diagnosis [16, 17]. Convolutional Neural Networks (CNNs) attain up to 97.8% accuracy in AD prediction, while EfficientNet-b0 and EEG-based ML models demonstrate robust performance in early-stage detection [18–20]. Explainable AI (XAI) techniques, such as SHAP values, provide interpretable insights into AD progression, with models like XGBoost achieving an F1-score of 0.84 by integrating genetic and clinical data [21, 24]. However, challenges persist, including small/imbalanced datasets and the need for clinically interpretable models [22, 23].

2.1 Limitations of Existing Approaches

Despite the progress made in medical science, accurately predicting AD remains challenging due to the unavailability of diverse datasets. Literature studies use MRI, PET, and CT scan images that are computationally expensive to handle and lack data diversity. This research aims to solve the problem of data diversity by utilizing two numeric datasets and also covers a broader age range from 33 to 96-year-old patients. This research uses DL techniques like Feed Forward Neural Network, Multi-layer Perceptron, Convolutional Neural Network, and ML techniques like Random Forest, Decision Tree, and Support Vector Machine that can accurately diagnose AD. However, we acknowledge potential computational challenges, emphasizing the need to balance accuracy and efficiency in practical implementation.

3. Research Methodology

In this study, two datasets named OASIS-1 and OASIS-2 available in Kaggle are deployed to identify AD. OASIS-1 includes cross-sectional T1-weighted MRI scans of 416 patients aged 18–96 years, whereas OASIS-2 provides longitudinal T1-weighted MRI scans of 150 patients aged 60–96 years. 226 records are taken from each dataset. To deal with missing values, mode and mean imputation procedures are used, for the case of the missing Social SES and the MMSE scores. New features are added, and the tables are cleaned by removing useless columns to

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

perform feature reduction, and the gender and handedness fields are encoded into binary form. The EDUC attribute values are then standardized by use of the Hollingshead Index of Social Position and the use of the z-score to assist in testing for outliers.

Feature ranking methods, such as k-best score, are applied to enhance the predictive models, followed by feature scaling using a standard scaler. The data is randomly split into training, validation, and testing sets, with k-fold cross-validation used for evaluation.

The study employs several machine learning classifiers to predict Alzheimer's Disease, including FFNN, MLP, CNN, SVC, RF, DT, and Voting Classifiers. The performance of these models is evaluated using metrics such as F1-Score, Recall, Accuracy, Precision, and AUC/ROC to ensure accuracy, reliability, and efficiency in detecting AD. The details of each step are illustrated in Figure 3.1.

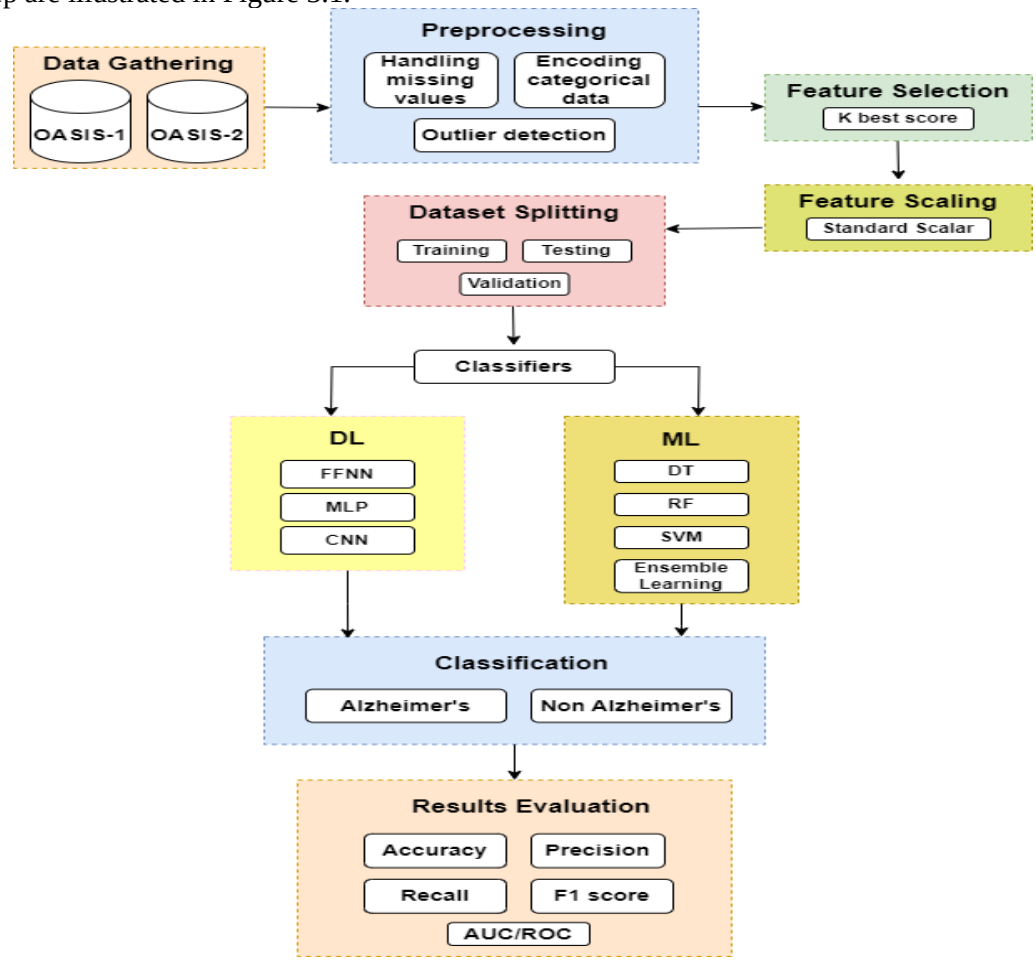


Figure 3.1: Framework Diagram of Alzheimer's Disease Detection using MRI data

3.1 Preprocessing

In Figure 1, Preprocessing refers to the set of activities and techniques applied to raw data before it is utilized in a machine learning or data analysis process, to improve data quality, resolve inconsistencies, and enhance the performance of the model or analysis.

3.2 Dataset

In this study, OASIS-1 and OASIS-2 publicly accessible datasets obtained from the Kaggle database having the MRI studies held at Washington University Alzheimer Disease Research

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Center (ADRC) are employed to detect AD consisting of cross-sectional MRIs of 416 patients aged 18 - 96 and OASIS-2 consists of longitudinal MRIs of 150 patients aged 60 – 96, and each participant undergoes at least one scan during the research. All participants are right-handed. This research takes an equal number of samples per gender (Male/Female) to eliminate potential biases [42]. 226 subject records are chosen from each OASIS-1 and OASIS-2. A total of 452 records is selected for this study.

Table 3.1 Dataset Description

Sr.No.	Attributes	Description
1	Subject ID	Patient ID
2	MR ID	MRI No of patients
3	Visit	Patients visit No
4	M/F	Gender: M for Male, F for Female
5	Hand	Left/Right
6	Age	Age of patients in years
7	EDUC	5 levels of Education (1=low to 5=high)
8	SES	Socio-Economic Status
9	MMSE	Mini-Mental State Examination (max score 30)
10	CDR	Clinical Dementia Rating (0 to 2)
11	eTIV	Estimated Total Intracranial Volume (cm^3)
12	nWBV	Normalize Whole Brain Volume calculated as a % of all voxels in the atlas-masked
13	ASF	Atlas Scaling Factor (unitless)
14	Delay	MRI Delay
15	Group	Alzheimer's and non-Alzheimer's

3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial initial step in data science projects that involves analyzing and visualizing data to understand its key characteristics, uncover patterns, and identify relationships between variables.

Table 3.2 Statistical Analysis of each attribute.

Feature	Min	Max	Mean	Median
Age	33	98	75.15	76
Educ	1	5	3.19	3
SES	1	5	2.55	3
MMSE	4	30	26.60	28
CDR	0	2	0.36	0.5
eTIV	1143	2004	1498	1476.5
nWBV	0.64	0.84	0.73	0.7295
ASF	0.87	1.53	1.18	1.1885

Tables 3.1 and 3.2 show all the variables' descriptions along with minimum, maximum, mean, and median values.

3.4 Handling Missing Values

Handling missing values is crucial in machine learning and data analysis for several key reasons. Firstly, missing data can significantly reduce the sample size, leading to biased and unreliable results, particularly with small datasets. Improper handling of missing values can also introduce bias, limit the types of analyses that can be performed, and negatively impact model performance. Addressing missing data helps maintain the integrity of the dataset,

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

enhances overall data quality, and enables more accurate descriptive statistics. Additionally, efficiently handling missing values can save time and effort during analysis, while ensuring a more unbiased representation of the underlying patterns in the data. By addressing missing data effectively, models can be trained on a more complete dataset, leading to improved performance, accuracy, and informed decision-making based on reliable, data-driven insights. Figure 3.2 shows the columns that have missing values. SES Socioeconomic status has 38, MMSE has 2, and MR Delay has 192 missing values.

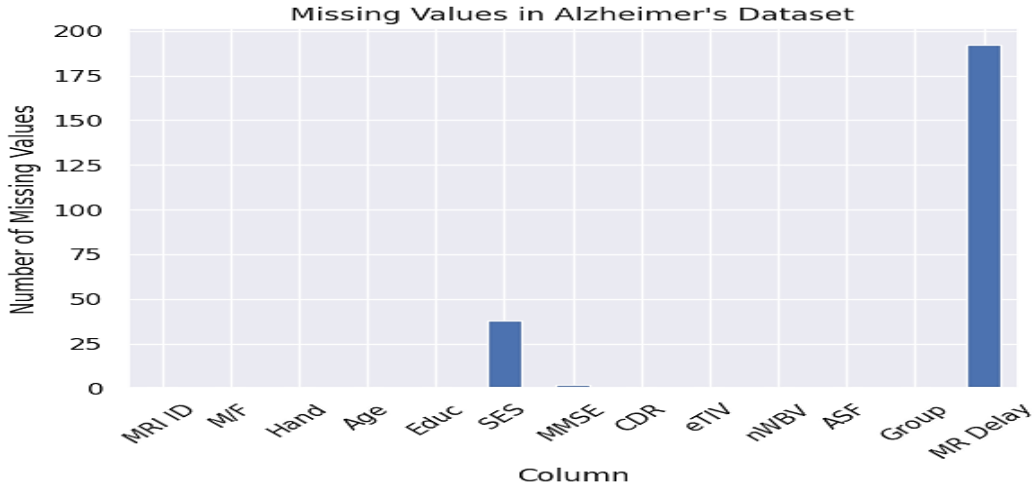


Figure 3.1 Missing values in the Alzheimer's dataset

In this research, we impute SES, which is a categorical attribute with the most frequent value (mode), and MMSE with the median.

3.5 Encoding Data

This research work uses label encoding, where each unique category is replaced with a unique numerical value. This is a common technique used to convert categorical data into numerical data that can be processed by machine learning algorithms. The categorical values 'R' and 'L' in the 'Hand' column are encoded with numerical values 1 and 0, respectively. This means that 'R' is encoded as 1 and 'L' is encoded as 0. The categorical values 'M' and 'F' in the 'M/F' column are encoded with numerical values 0 and 1, respectively. This means that 'M' is encoded as 0 and 'F' is encoded as 1. Similarly, the categorical values 'Alzheimer's' and 'Non-Alzheimer's' in the 'Group' column with numerical values 1 and 0, respectively. This means that 'Alzheimer's' is encoded as 1 and 'non-Alzheimer's' is encoded as 0.

3.6 Outlier Detection

Outlier identification is an important process that happens during data pre-processing and data analysis, as it aids in the detection of significant figures that deviate from the remaining data. Since there are different techniques for detecting outliers, one of the commonly used tests is the statistical one, such as the Z-score, which measures the distance of a point from the mean. Another commonly used test is the Interquartile Range (IQR), which identifies points that are outside the IQR range. This study utilizes the z-score method to detect outliers. It determines the bounds for outliers and collects the indices of rows where the values fall outside these bounds.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

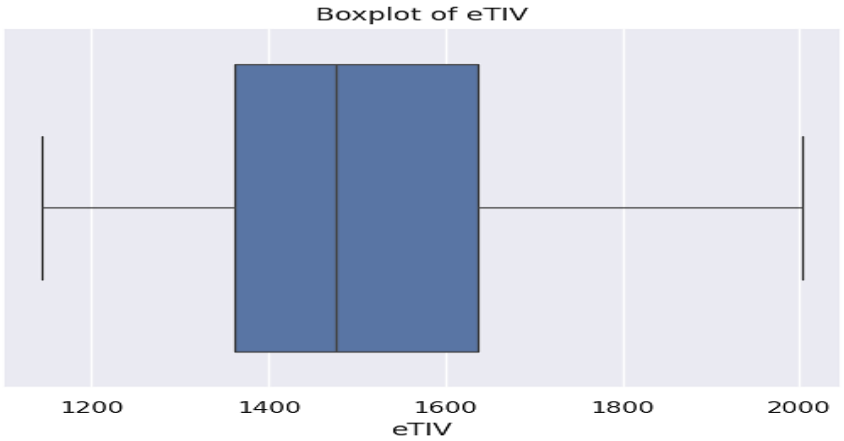


Figure 3.2 Boxplot of eTIV to detect outliers

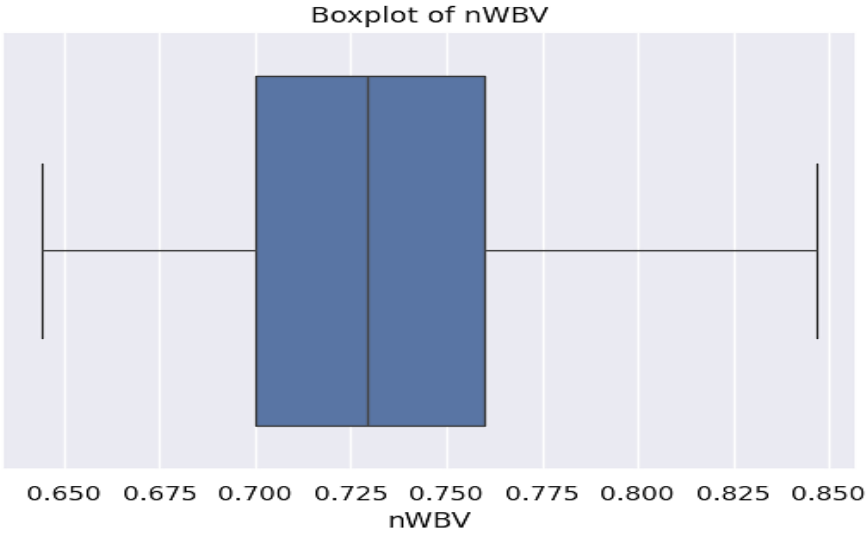


Figure 3.3 Boxplot of nWBV for outlier detection

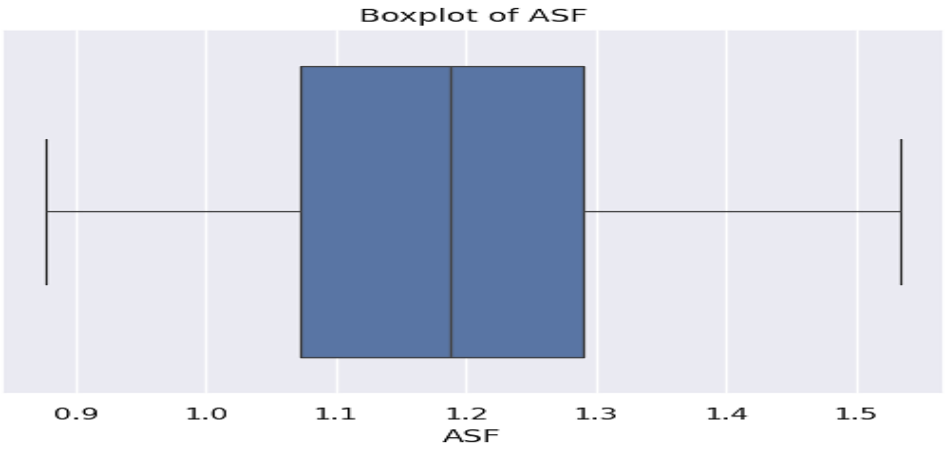


Figure 3.4 Boxplot of ASF for outlier detection

Figures 3.3, 3.4, and 3.5 show the boxplot of the three numerical features. The boxplot shows that there are no outliers in these three attributes.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

3.7 Feature Selection:

Feature selection is an important step in implementing a machine learning algorithm with pre-processed data that involves determining which variables to use for constructing an effective model. This technique minimizes the data representation in lower-dimensional feature space by removing both redundant and potentially noisy features while maintaining a nearly optimal decision function of the trained model. Feature selection is an essential task since reducing the number of feature inputs has a positive effect on the accuracy of the model when tested on new data and in preventing overfitting. Furthermore, it helps to improve the interpretability of the model by selecting which variables to use in the next model, making it easier to understand the model's results and the relationships between the variables. This study uses the correlation coefficient and score function for feature ranking.

3.7.1 Correlation Coefficient:

It is a matrix where the values represent correlation coefficients between variables in a given data stream. It is usually an n x n square matrix where the values in the matrix are the correlation between variables. The Correlation Matrix consists of the correlation between all those variables, which makes it easy to judge the strong correlation of variables and other relations as well.

3.7.2 Score Function-Based Feature Selection

In the score function-based feature selection methods, statistical measures are applied to determine the suitability of each of the features to the targeted variable. These methods decide a score for each feature as per their relation or dependency with the results or the variable under consideration. The features are then sorted, and the ones with the highest scores are chosen as decision-makers.

Figure 3.6 shows feature ranking using the score function, in which CDR got the highest rank and M/F got the lowest rank.

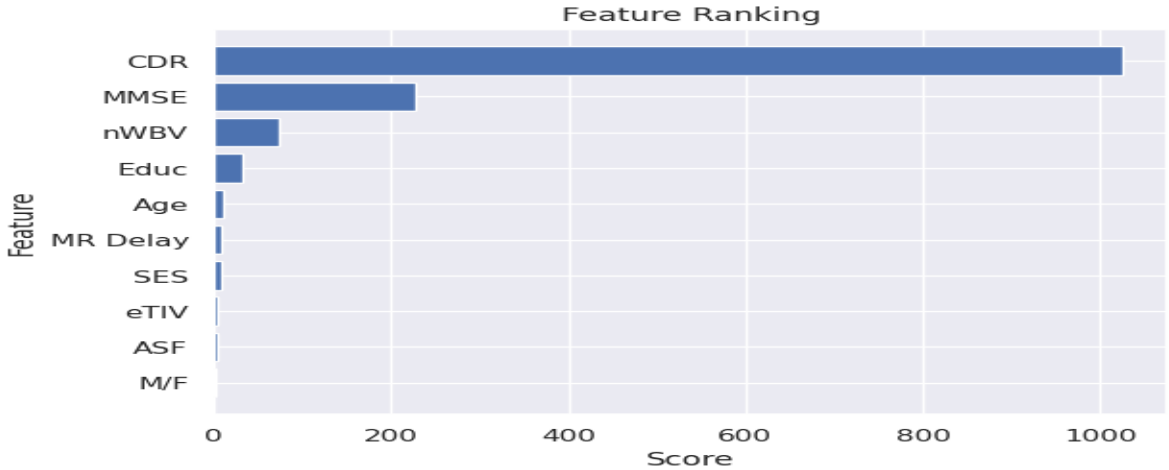


Figure 3.5 Feature Ranking using the score function

3.7.3 Heat Map:

A heatmap can be described as a graphical display of a set of data presented in a correlation matrix. It is an obvious grid of the correlation coefficients in which the rows and the columns denote the variables, whereas the cells indicate the correlation nexus. The heatmap then

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

displays values as colors so that one can distinguish the pattern and the pattern of the direction of the correlation.
Figure 3.7 heatmap shows the correlation of all the attributes with the target class Group.

Strong correlations (absolute value > 0.5):
nWBV: Strong positive correlation with "Group", indicating that higher nWBV values are associated with higher values in the "Group" variable.
ASF: Strong positive correlation with "Group", suggesting that higher ASF values are associated with higher values in the "Group" variable.
eTIV: Strong negative correlation with "Group", indicating that higher eTIV values are associated with lower values in the "Group" variable.

Weak correlations (absolute value < 0.25):
M/F: Weak correlation with "Group", suggesting that the male/female feature has little influence on the "Group" variable.
Age: Weak correlation with "Group", indicating that age has a limited impact on the "Group" variable.
Educ: Weak correlation with "Group", suggesting that education level has a relatively small influence on the "Group" variable.
SES: Weak correlation with "Group", indicating that socioeconomic status has a limited impact on the "Group" variable.
MMSE: Weak correlation with "Group", suggesting that the Mini-Mental State Examination score has a relatively small influence on the "Group" variable.
CDR: Weak correlation with "Group", indicating that the Clinical Dementia Rating has a limited impact on the "Group" variable.
MR Delay: Weak correlation with "Group", suggesting that the MR Delay feature has a relatively small influence on the "Group" variable.

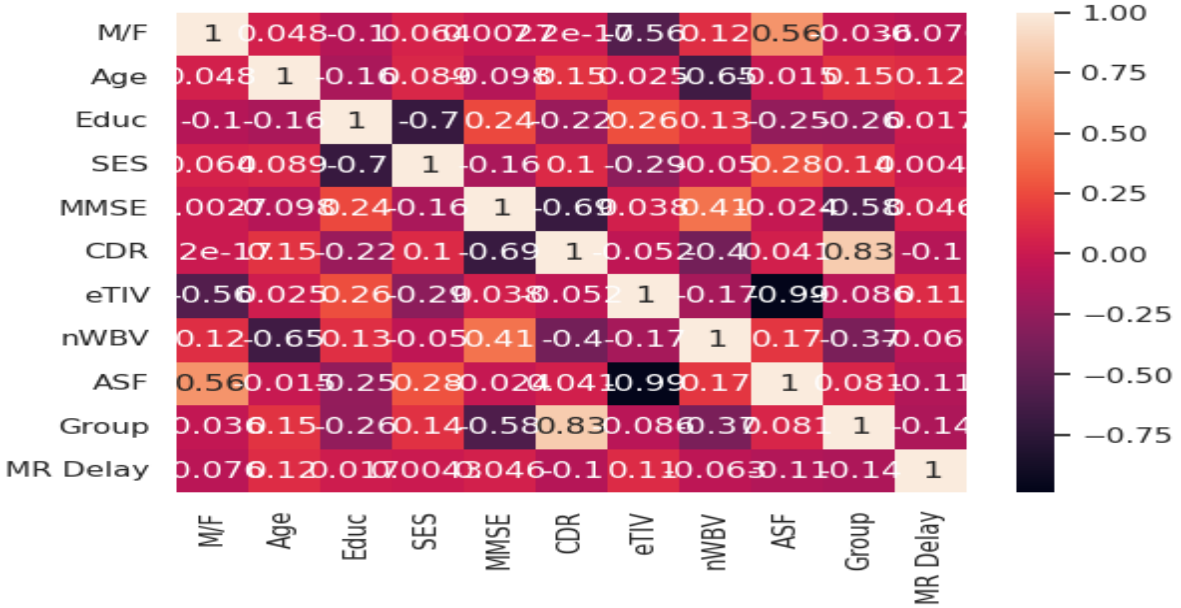


Figure 3.6 Heat map showing the relation between each attribute

3.8 Data Splitting

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Training-Testing Split In machine learning, data is divided into various subsets for different uses; it can, for instance, be used to train an algorithm and test its performance on unknown data. Splitting data is a very important aspect when it comes to testing models so that one can get a true measurement of the model as well as solve each of the overfitting and underfitting problems. The typical split involves dividing the dataset into three main subsets: pre-training, validation, and test sets or data for further training and enhancing world knowledge.

The dataset is usually divided into specific subsets of data to train the machine learning algorithms, although the largest percentage is provided to the training set. In this configuration, 60%, 70%, 75%, and 80% of data subsets are used to perform analysis on which subset our model performs well. This significant fraction enables the model to learn many detailed aspects inherent in the data, such as patterns, interdependence, and aspects.

Another important and relatively less extensive dataset in machine learning is the validation or verification set. Comprising around 10-20% of the entire data population, this niche population is immensely essential for enhancing the degree of accuracy in your models. The validation set is a tool to tune hyperparameters and check the adaptation of the model. This step avoids overfitting.

There is also another and rather separate category that is unseen during the training or validation of the machine learning model and is meant for testing purposes only within the framework of the given approach. This constitutes roughly 20%, 25%, 30% and 40% of the data subsets that are used for testing accuracy. Regarding the division of our data, we are using K-fold cross-validation. Huge potential in data splitting for constructing accurate and generalizable machine learning models while giving a true estimation of their performance in totally unseen data.

3.9 Classifier

In our methodology, we adopt a comprehensive approach for binary classification by integrating Deep Learning algorithms like Convolutional Neural Networks, Feed-Forward Neural Networks, and Multi-Layer Perceptron. Machine Learning algorithms like Decision Trees, Random Forests, and Support Vector Machines. An ensemble Learning Technique, like a Voting classifier, is used. This integrated approach ensures that our methodology leverages the strengths of each component, providing a robust solution for accurate and interpretable binary classification.

3.9.1 Support Vector Machine (SVM):

It is an easily understandable algorithm that falls under the category of Supervised learning and is used for both classification and regression. The major concept is to identify a hyperplane in a large number of dimensions, which independently and maximally separates data points of different classes, defined as the support vectors lying on the sidelines of the boundary. During learning, SVM utilizes limited labeled data to find the right hyperplane that best separates the two forms of classes in the data sets: experiential data and spectral data, modelling a true target value in predictor variables for test data.

3.9.2 Random Forest

Random Forest is an advanced type of ensemble learning approach that uses a range of decision trees within a single algorithm to bring greater prediction accuracy and minimal mis forecasting. It operates by using each tree to learn with some arbitrary features and instances, and after that, combining their predictions to offer the final result. It also improves interpretability and can deal with a high number of features; thus, Random Forests are widely used when aiming at classification and regression.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

3.9.3 Decision Tree

Decision Trees are one of the branches of Supervised Machine Learning, mainly used in classification and regression models. It divides the feature space into smaller subspaces depending on further information gained about the features and structures the subspace by a tree-like structure of the decision and its outcome. Another advantage of Decision Trees is that they give an understandable and clear understanding of the decision-making process, they are also easy to interpret and use, and both numerical and categorical data can be used to make decisions.

3.9.4 Multi-Layer Perceptron

MLP is another type of feedforward ANN that has an input layer, one or more hidden layers, and an output layer. In the learning process, MLP utilizes the algorithm of backpropagation; besides, by employing a single hidden layer, they are capable of approximating any measurable function, which contributes to the applicability of these networks in such tasks as classification, regression, and pattern recognition. Because of this, it is widely used for its superior performance in learning nonlinear mappings from inputs in the data to appropriate outputs.

3.9.5 Convolutional Neural Network

They are known as deep learning models or more specifically, CNNs designed for structured grid data, such as images or video frames, and also for numeric datasets. They employ convolutional layers for the hierarchical features extraction, such that they use pooling layers for dimensionality reduction and fully connected layers for the classification. CNN performs fairly well in areas such as image recognition, object detection, image segmentation, and classification tasks, especially considering its ability to capture the hierarchical spatial structure of data.

3.9.6 Feed Forward Neural Network

FFNN is a type of artificial neural network where the connections or paths formed between the nodes do not have a feedback loop and have only a one-way flow of information from input to output. The interconnected structure of FFNN includes the input layer, one or more hidden layers, and the output layer. They are mainly used in supervised learning when we use them in classification and regression functions, and it is proven that they can approximate real-valued functions arbitrarily well.

3.9.7 Voting Classifier

A classification task in which an individual classifier combines more than one prediction from other individual classifiers for deriving a final decision is known as an ensemble learning technique. Through the usage of these ML and DL classifiers, we propose to harness the best from the two to reach enhanced precision in the identification of Alzheimer's in the early stages.

3.10 Performance Evaluation Measures

In this part, we have covered the evaluation measure that will be used in this study to assess the model's effectiveness.

Accuracy: In deep learning, the accuracy score is a metric that contrasts the percentage of correctly predicted data made by a model with the total number of predictions. By dividing the total number of forecasts by the number of accurate estimates, we can calculate it using equation I below.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Precision: Precision, or the Caliber of a successful prediction the model produces, is one measure of the model's effectiveness. To calculate accuracy, divide the total number of positive predictions by the sum of true positives and false positives, or the percentage of true positives. We can calculate it using equation II below.

.

$$precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall: The percentage of correctly recognized Positive samples among all Positive samples is used to calculate the recall. Recall measures the model's ability to recognize positive samples. The recall increases with the number of positive samples found; we can calculate it using equation III below.

.

$$recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1 score: The F1 score is one of the most important evaluation metrics in machine learning. It summarizes a model's capacity for prediction elegantly by combining recall and accuracy, two metrics that would typically be in opposition to one another. We can calculate it using equation IV below.

.

$$f1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

AUC ROC: The Area Under the Receiver Operating Characteristic Curve is a performance metric used to check the ability of a binary classification model to differentiate between positive and negative classes across various threshold settings.

4. Experimental Results

In the context of prenatal care, this thesis focuses on advancing methodologies for Alzheimer's classification using ML and DL models. The research explores the integration of these models to analyze numeric factors derived from the dataset related to Alzheimer's. The initial phase involves comprehensive data visualization and analysis to understand the characteristics and complexities of the dataset used in this study.

Building upon this groundwork, the thesis presents results derived from the proposed methodology, which incorporates advanced Deep and Machine Learning techniques, while voting is employed as the final classification method to enhance model performance and accuracy in detecting Alzheimer-related factors.

The quantitative evaluation of the model's performance focuses on its discriminative capabilities in detecting and assessing factors associated with Alzheimer's within the datasets. Through this structured approach, the research aims to contribute significantly to improving diagnostic capabilities for Alzheimer's, thereby enhancing mental health.

4.1 Data Visualization

Figure 4.1 shows data visualization representing the distribution of patients across each class in our dataset. Each bar in the chart corresponds to a class, and the height of the bar indicates the number of records present in that class. This visualization provides insights into the composition of our dataset, helping us understand the diversity and balance of data among each class. Our dataset dimension is (452 X 13).

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

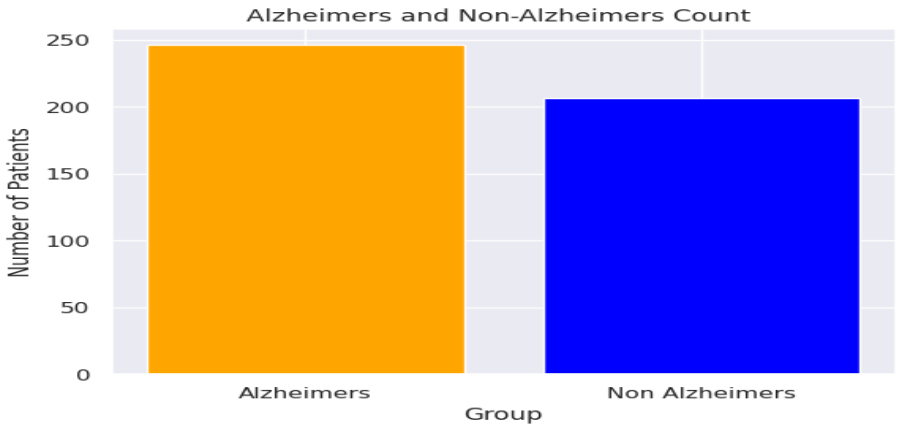


Figure 4.1 Data distribution in each class

Figure 4.1 illustrates the distribution of individuals in the dataset based on the presence or absence of Alzheimer's. Specifically, it shows that there are 246 records corresponding to individuals diagnosed with Alzheimer's and 206 records corresponding to individuals not affected by Alzheimer's. This distribution highlights the balance between the two classes within the dataset, with a relatively equal representation of individuals with and without Alzheimer's. Understanding this distribution is crucial for assessing the dataset's suitability for modeling and analysis. A balanced representation of classes helps prevent bias in predictive models and ensures that the classifier is trained effectively to distinguish between individuals with and without Alzheimer's. This insight gained from Figure 4.1 informs subsequent analyses aimed at developing accurate and reliable predictive models for Alzheimer's detection in health care.

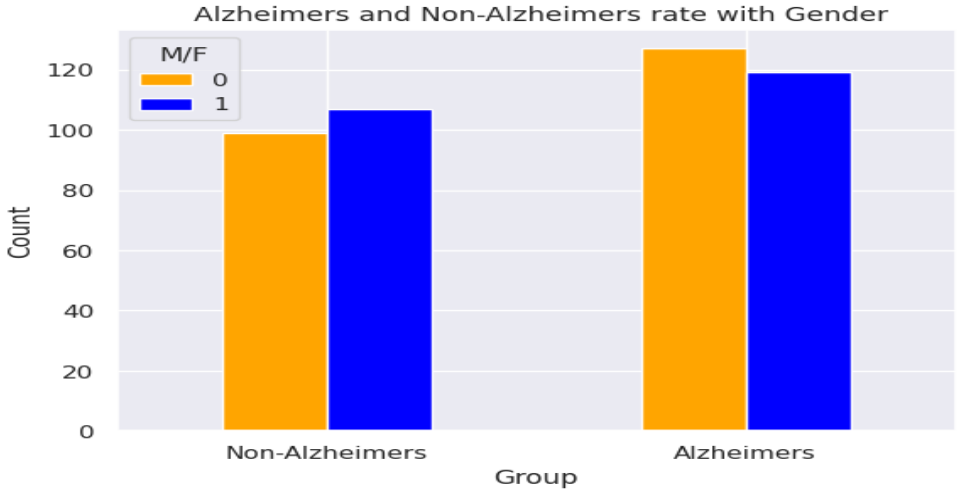


Figure 4.2 Males and Female patients in the Dataset

Figure 4.2 illustrates the distribution of males and females in the dataset based on the presence or absence of Alzheimer's. 127 males have Alzheimer's, and 99 are healthy. Similarly, 119 females have Alzheimer's, and 107 are healthy. This shows that men have a higher risk of Alzheimer's as compared to women.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

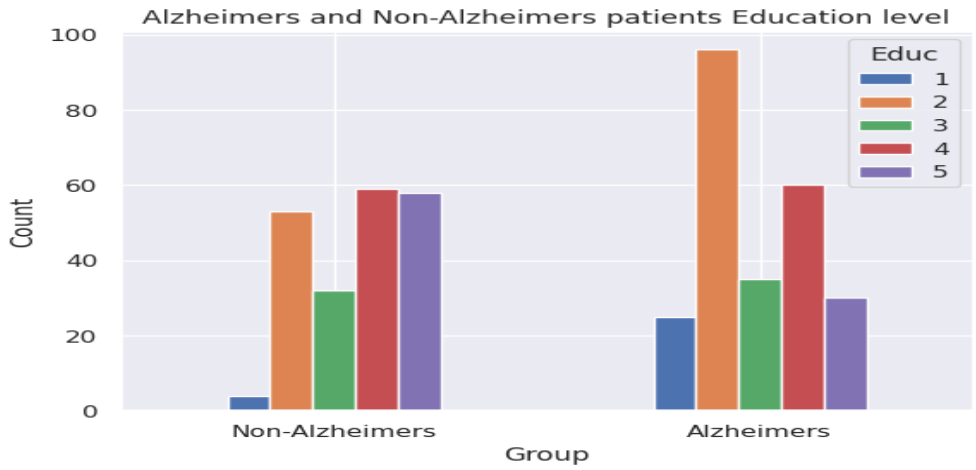


Figure 4.3 Education level of Alzheimer's and Non-Alzheimer's patients

Figure 4.3 illustrates the distribution of patients' education level in the dataset based on the presence or absence of Alzheimer's. Education codes correspond to the following 5 levels of education: 1: less than high school grad., 2: high school grad., 3: some college, 4: college grad., 5: beyond college. Figure 4.3 shows that those who have a high education level are less prone to Alzheimer's as compared to low education level.

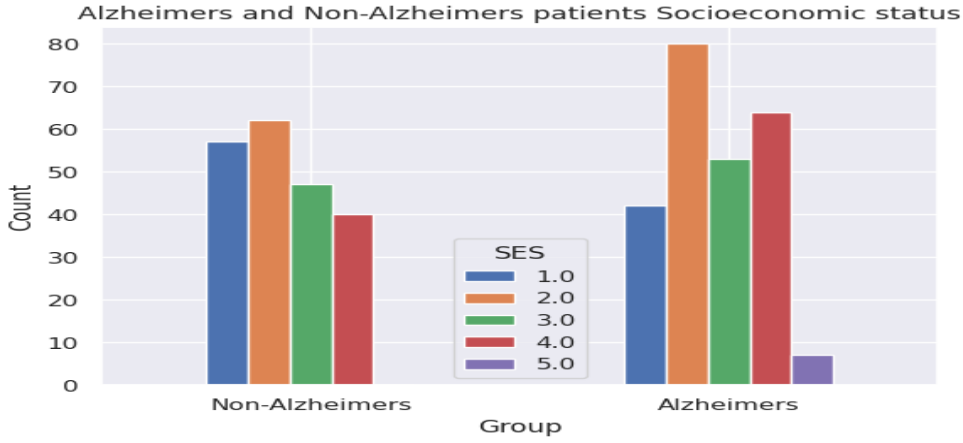


Figure 4.4 Socioeconomic status of patients

Figure 4.4 illustrates the distribution of patients' socioeconomic status (SES) according to the Hollingshead index of social position in the dataset based on the presence or absence of Alzheimer's. The Hollingshead Index of Social Position is a survey designed to measure an individual's social status based on four domains: marital status, retired/employed status, educational attainment, and occupational prestige. Figure 4.4 shows that individuals having high SES are less prone to Alzheimer's.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about



Figure 4.5 Clinical Dementia Rating of Patients

Figure 4.5 illustrates the distribution of patients' Clinical Dementia Rating (CDR 0= nondemented; 0.5 = very mild dementia; 1 = mild dementia; 2 = moderate dementia) in the dataset based on the presence or absence of Alzheimer's. All participants with dementia (CDR >0) were diagnosed with probable AD.

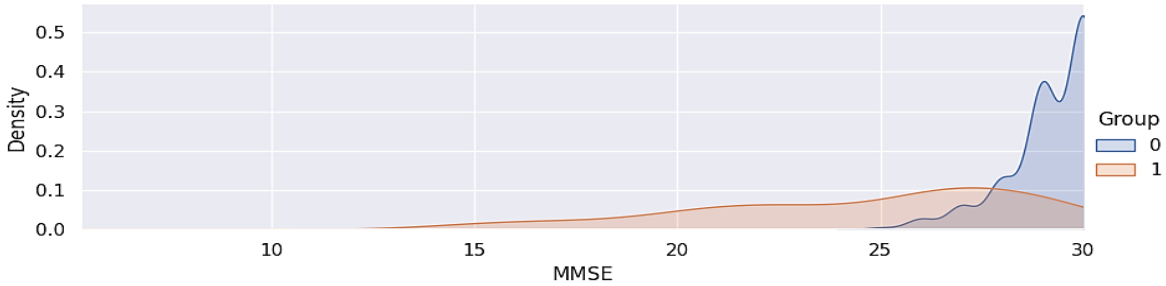


Figure 4.6 MMSE scores of patients

Figure 4.6 illustrates the Mini-Mental State Examination (MMSE) in the dataset based on the presence or absence of Alzheimer's. It is a 30-point test used to measure orientation, immediate memory, attention, delayed memory, and language. Its range is 0-30. A score of 25 or higher is classified as normal. Figure 4.6 shows that patients having a score of 24 and below indicate cognitive impairment.

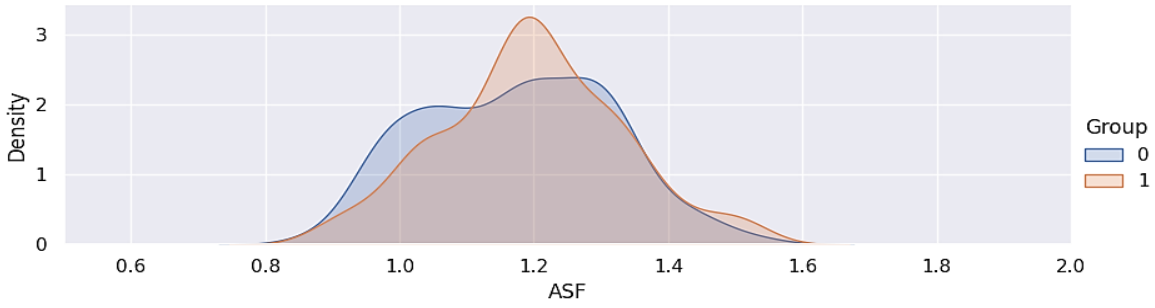


Figure 4.7 ASF scores of patients

Figure 4.7 illustrates Atlas Scaling Factor (ASF) is a useful biomarker derived from brain MRI that reflects brain atrophy. it is a scaling factor that normalizes the brain volume to a standard

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

atlas. A lower ASF indicates a smaller brain size compared to the atlas.

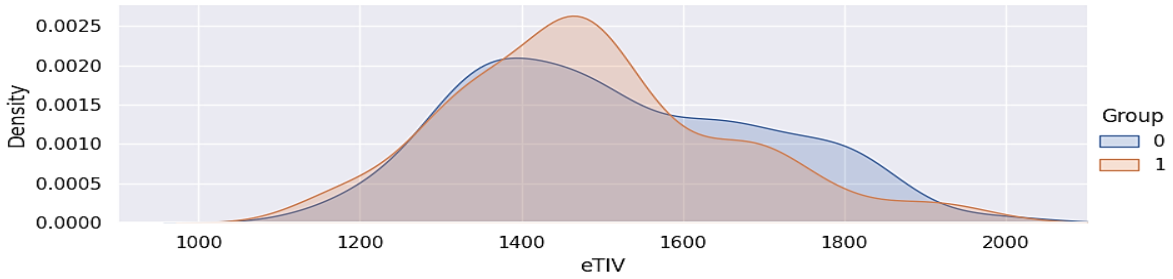


Figure 4.8 Estimated total intracranial volume (eTIV) (mm3) of patients

Figure 4.8 illustrates estimated total intracranial volume (eTIV) is an important MRI-derived measure that reflects the volume of the intracranial cavity, which includes the brain, cerebrospinal fluid, and other tissues. A lower eTIV can be an indicator of a smaller brain size.

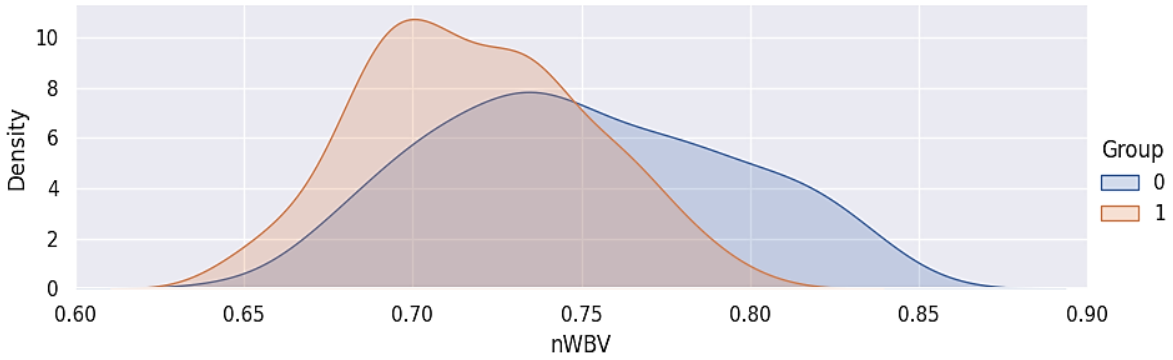


Figure 4.9 Normalized whole-brain volume (nWBV) of patients

Figure 4.9 illustrates that normalized whole-brain volume (nWBV) is an important MRI-derived measure that reflects the percentage of the intracranial cavity occupied by the brain. A lower nWBV indicates a smaller brain size.

Figures 4.7, 4.8, and 4.9 show that Alzheimer's patients have a higher brain volume ratio than the Alzheimer's group. The reason for this is that the diseases influence the brain tissues, causing them to shrink.

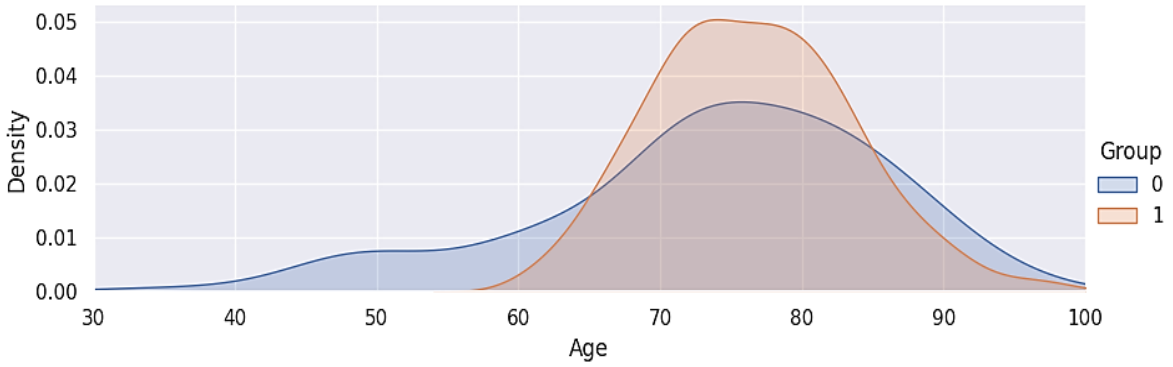


Figure 4.10 Age of patients in the Dataset

Figure 4.10 illustrates the age of patients in the dataset based on the presence or absence of Alzheimer's. Figure 4.10 shows that People above the age of 60 are more prone to Alzheimer's.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

4.2 Hyperparameters

A hyperparameter is a parameter that is set before training a machine learning or deep learning model, and it controls how the learning process works. They act as "settings or knobs" that guide the model training process. Correct tuning ensures the model is neither too simple (underfitting) nor too complex (overfitting), leading to optimal performance.

To ensure fair comparison, the hyperparameters of all ML models were systematically tuned. Table 4.1 summarizes the selected hyperparameters and their optimal configurations.

Table 4.1: Hyperparameters of Machine Learning Models used for Alzheimer's Classification

ML Models	Hyperparameters	Optimal Values
Decision Tree (DT)	Random State, Max Depth, Criterion	0, 10, Gini
Random Forest (RF)	n_estimators, max_features, max_depth	14, 8, 8 (best in range)
Support Vector Machine (SVM)	C, Gamma, Kernel	C = 100, Gamma = 10, Kernel = RBF
Ensemble (Voting)	Type	Soft Voting (94.50%), Hard Voting (90.10%)

Table 4.1 shows the tuned hyperparameters for ML models. Among them, the ensemble soft voting method achieved the highest accuracy (94.50%), outperforming individual classifiers. Deep learning (DL) models were trained with different architectures and hyperparameter settings to capture complex feature representations. Table 4.2 presents the configuration of CNN, MLP, and FFNN models.

Table 4.2: Hyperparameters of Deep Learning Models used for Alzheimer's Classification

DL Models	Hyperparameters	Values/Configuration
CNN	Filters, Kernel Size, Pooling, Activation, Optimizer, Loss, Epochs	32–64 filters, kernel = 2, MaxPooling, ReLU, Adam, Binary Crossentropy, 100 epochs
MLP	Optimizer, Loss, Batch Size, Dropout, Validation Split, Epochs	RMSprop, Binary Crossentropy, Batch = 32, Dropout = 0.5, Val Split = 0.2, 100 epochs
FFNN	Optimizer, Loss, Batch Size, Epochs	Adam, Binary Crossentropy, Default batch, 100 epochs

Table 4.2 summarizes the hyperparameters of DL models. CNN achieved the most balanced performance, while MLP and FFNN also showed competitive results, demonstrating the effectiveness of deep architectures for Alzheimer's classification.

4.3 Result Evaluation

Tables 4.3, 4.4, 4.5, and 4.6 display the performance measures of the resulting model for Binary-class classification on the 4 training and testing sets. The most used measurements from the details list include Precision, Recall, F1 score, and Accuracy.

Table 4.3 Classification result and accuracy of each model, 60% training and 40% testing

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Model	Precision %	Recall %	F1 score %	AUC %	Accuracy %
DT	82.79	81.91	82.35	81.76	81.76
RF	88.88	**85.10**	86.95	86.80	86.74
SVM	91.66	81.91	86.51	86.93	86.74
Soft Voting	86.66	82.97	84.78	93.71	84.53
Hard Voting	87.77	84.04	85.86		85.63
CNN	91.76	82.97	87.15	94.17	87.29
MLP	**93.58**	77.65	84.88	94.44	85.63
FFNN	91.86	84.04	**87.77**	**94.63**	**87.84**

Table 4.4 Classification result and accuracy of each model, 75% training and 25% testing

Model	Precision %	Recall %	F1 score %	AUC %	Accuracy %
DT	89.79	73.33	81.08	81.94	81.41
RF	**97.95**	80.00	88.07	89.05	**88.49**
SVM	88.88	80.00	84.21	84.33	84.07
Soft Voting	90.56	80.00	84.95	92.86	84.95
Hard Voting	94.11	80.00	86.48		86.72
CNN	90.00	**90.00**	**90.00**	**94.15**	**89.38**
MLP	**93.61**	73.33	82.22	93.05	83.18
FFNN	90.74	81.66	85.96	92.07	85.84

Table 4.5 Classification result and accuracy of each model, 70% training and 30% testing

Model	Precision %	Recall %	F1 score %	AUC %	Accuracy %
DT	82.60	81.42	82.14	81.62	81.61
RF	89.70	87.14	88.40	88.26	88.23
SVM	**94.11**	91.42	**92.75**	92.68	**92.64**
Soft Voting	91.17	88.57	89.85	**96.86**	89.70
Hard Voting	85.29	82.85	84.05	-	83.82
CNN	87.67	**91.42**	89.51	94.39	88.97
MLP	93.22	78.57	85.27	93.85	86.02
FFNN	91.93	81.42	86.36	94.35	**86.36**

Table 4.6 Classification result and accuracy of each model, 80% training and 20% testing

Model	Precision %	Recall %	F1 score %	AUC %	Accuracy %
DT	97.56	78.43	86.95	87.96	86.81
RF	95.45	82.35	88.42	88.67	87.91
SVM	94.11	94.11	94.11	93.30	93.40
Soft Voting	96.00	94.11	95.04	98.43	**94.50**
Hard Voting	97.72	84.31	90.52	-	90.10
CNN	95.34	80.39	87.23	94.80	86.81
MLP	87.23	80.39	83.67	92.89	82.41
FFNN	90.00	88.23	89.10	93.38	**87.91**

The evaluation tables give a detailed description of the model's performance for binary classes. The classification results for various models evaluated on different training and testing splits reveal significant insights into their performance. When trained and tested on 60% and 40%

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

splits, the FFNN model achieved the highest accuracy at 87.84%, closely followed by the CNN model at 87.29%. As the training data increased to 75%, the RF model stood out with an 88.49% accuracy, with the CNN model maintaining high performance at 89.38%. For the 70% training split, the SVM model excelled with a 92.64% accuracy, while Soft Voting achieved the highest AUC at 96.86%. Finally, with 80% training data, the SVM model demonstrated superior performance with a 93.40% accuracy, and Soft Voting again had the highest AUC at 98.43%. Overall, the FFNN and CNN models consistently performed well across various splits, with Soft Voting and SVM showing remarkable performance in higher training data scenarios.

4.4 Confusion Matrix

The Confusion matrix is a measurement tool that can be used to assess a classifier because it provides a matrix of errors that separates the actual and recorded ones. This is one of the most effective measures or metrics in machine learning as well as in deep learning to compare the quality of the particular target model with the required model by determining the accuracy, precision, and recall, along with the F1 score of the particular model. A confusion matrix for a binary classification problem has a 2x2 structure:

Table 4.7 Confusion Matrix Components

	Predicted Negative	**Predicted Positive**
Actual Negative	True Negatives (TN)	False Positives (FP)
Actual Positive	False Negatives (FN)	True Positives (TP)

For multi-class problems, the confusion matrix is larger, with each row and column representing a class.

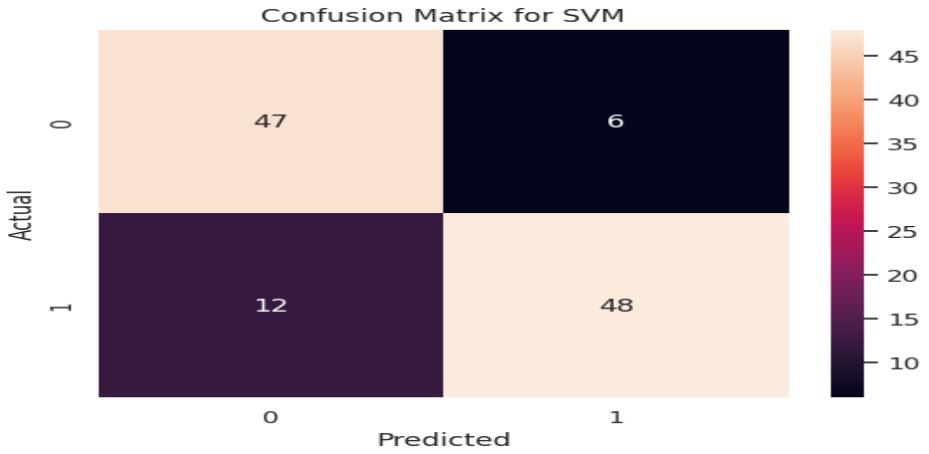


Figure 4.11 CM of SVM

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

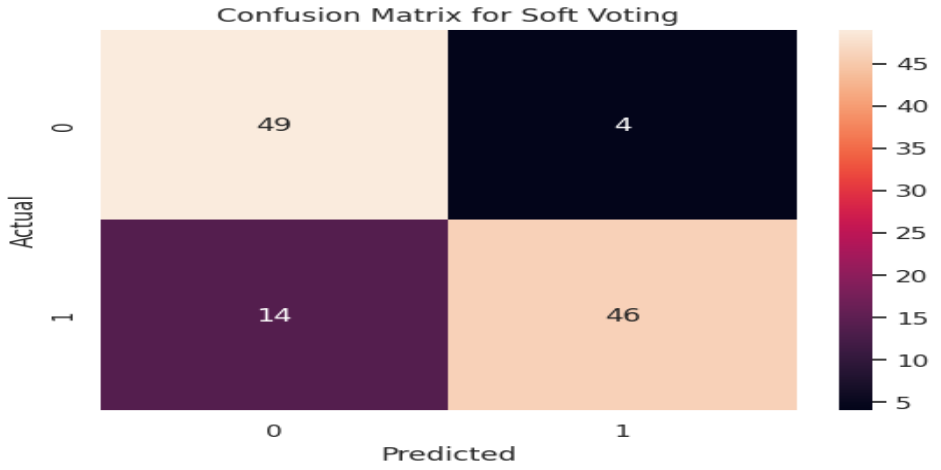


Figure 4.12 CM of Soft Voting

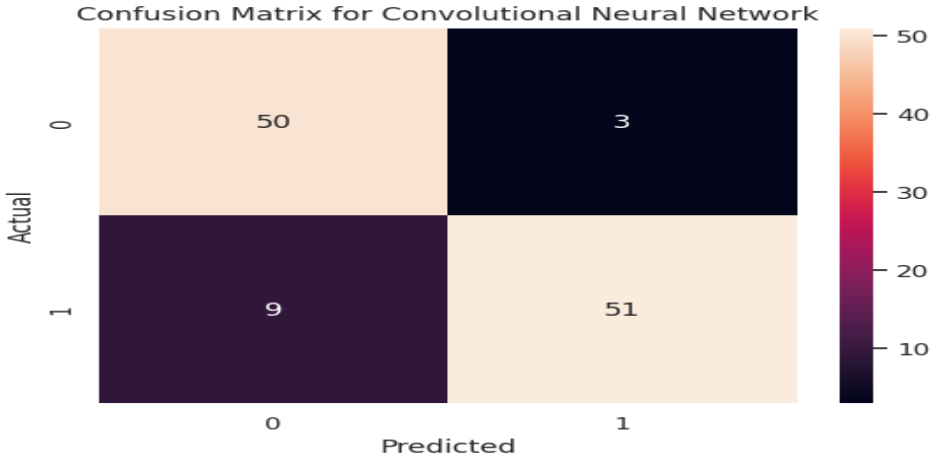


Figure 4.13 CM of CNN

Figures 4.11, 4.12, and 4.13 show Confusion matrices for ML and DL models having high accuracies.

ROC CURVE:

The Receiver Operating Characteristic graph is a graphical display of the true positive rates against the false positive rates, which is useful in assessing the performance of models that make a binary classification. It shows the performance of a classifier in terms of TPR and FPR graphs for different default values of thresholds.

True Positive Rate (TPR): It is also called Sensitivity/Recall, and this metric calculates the ratio of actual positive cases that a model is correctly flagging.

False Positive Rate (FPR): It computes the ratio of false positives to the total actual negatives, meaning how many out of all the points called negative are indeed negative by their nature.

Threshold: is the important and specific calibrating value defining the ratio, under which the probability score is to be recognized as positive. By changing the threshold, completely different TPR/FPR pairs can be observed. Below are the ROC curves for all the ML and DL models.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

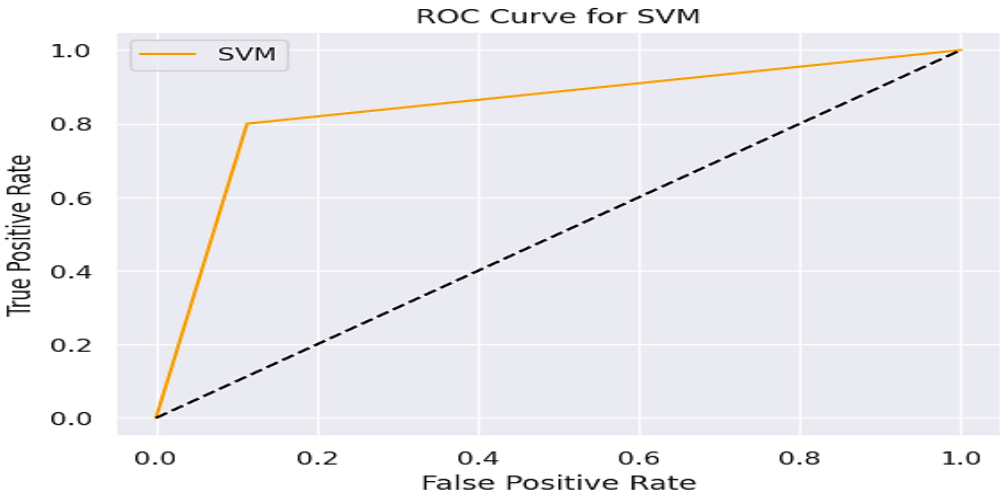


Figure 4.14 ROC Curve for SVM

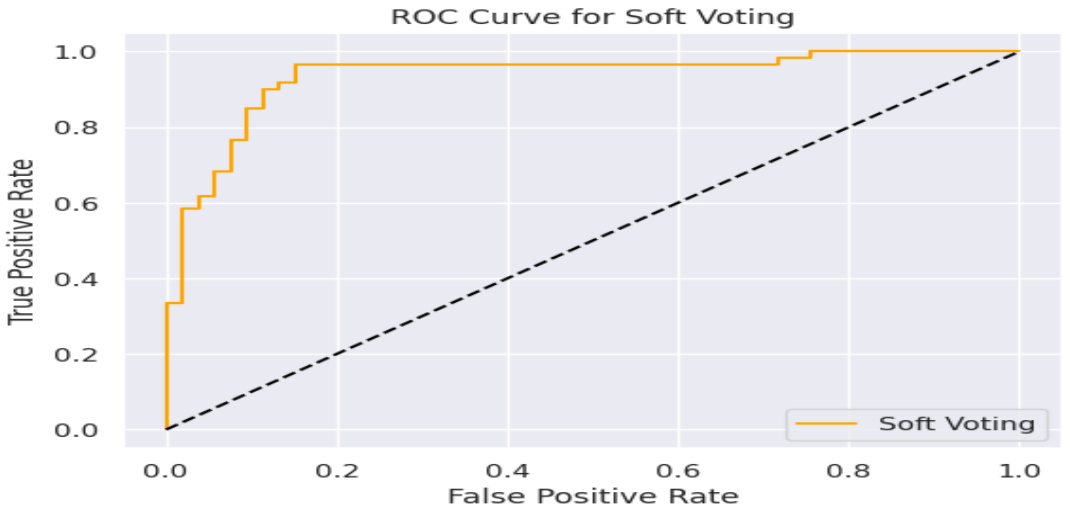


Figure 4.15 ROC Curve for Soft Voting

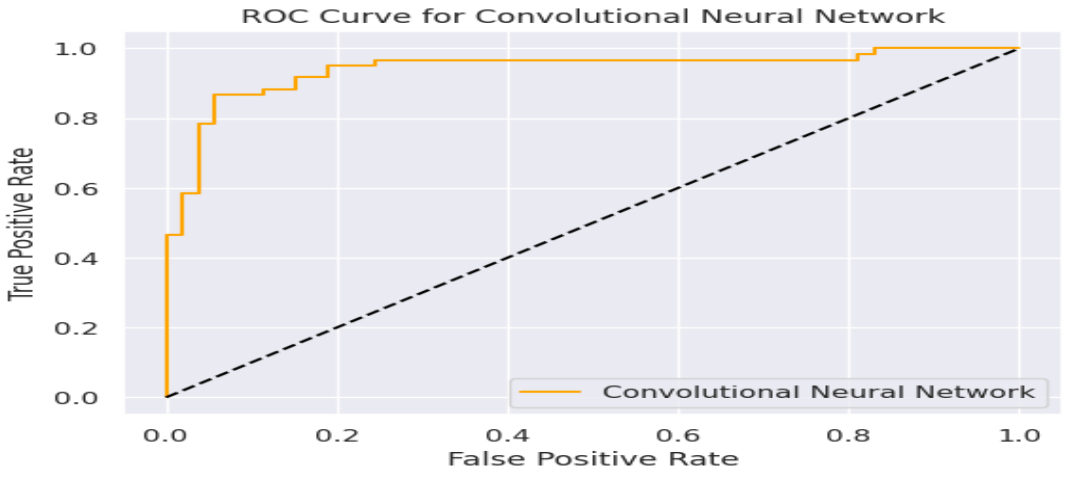


Figure 4.16 ROC Curve for CNN

Figures 4.14, 4.15, and 4.16 show ROC curves for ML and DL models having high accuracies.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

4.5 Comparison with existing techniques

Table 4.8 shows the comparison of our proposed method with the existing methods. FFNN achieved the highest accuracy of 87.84% with a 60% training split and maintained strong performance across other splits. CNN also showed excellent results, reaching up to 89.38% accuracy with a 75% training split and maintaining competitive performance with other splits. SVM demonstrated strong performance with an accuracy of 93.40% on an 80% training split, making it the best-performing machine learning model. RF also performed well, achieving up to 88.49% accuracy with a 75% training split.

The Soft Voting model has the highest accuracy of 94.50%, showing excellent discriminative ability.

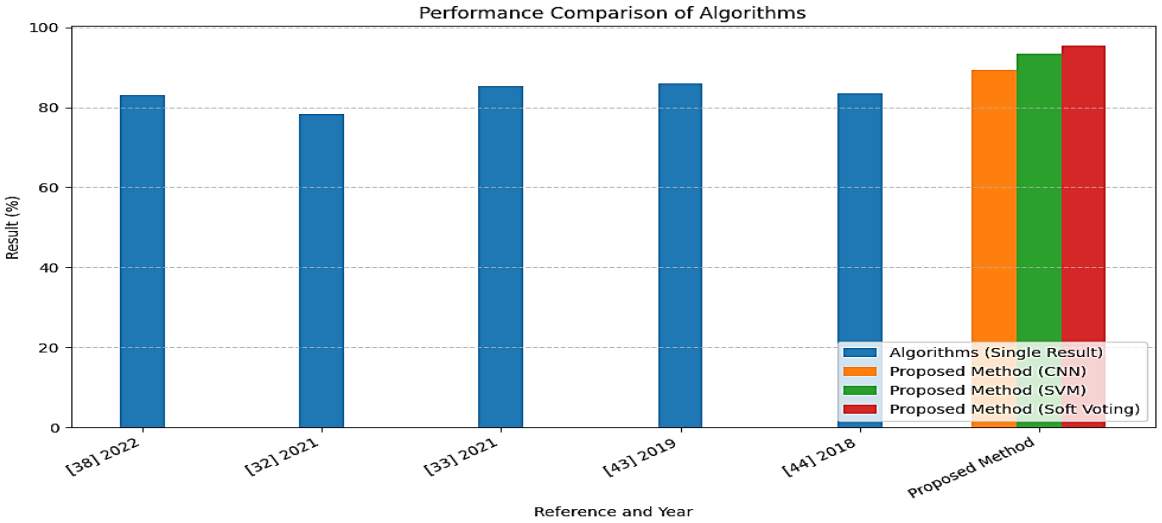


Figure 4.17 Comparison with existing Algorithms

Table 4.8 Comparison with existing techniques

Reference	Year	Algorithms	Result (%)
[38]	2022	DT, RF, SVM, Gradient Boosting, Voting classifier.	83%
[32]	2021	ML approaches (SVM) and PCA	78.31%
[33]	2021	Deep Neural Network	85.19%
[43]	2019	3D CNN	86%
[44]	2018	Cross-Modal Transfer Learning	83.57%
Proposed Method	2025	DT, RF, SVM, Voting classifier, FFNN, CNN, MLP	89.38% CNN 93.40% SVM 94.50% Soft Voting

5. Conclusion

Since at present, there is no cure for Alzheimer's disease, it is more crucial nowadays than ever to decrease risk, determine early stages, and evaluate the manifestations. The literature review proves that great research has been done to search for Alzheimer's disease through machine learning techniques and algorithms and micro-simulation approaches; however, it remains challenging to derive different features that can help to detect Alzheimer's at an initial stage.

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

For this study, various techniques are employed, namely, Decision Tree, Random Forest, Support Vector Machine, Convolutional Neural Network, Feedforward Neural Network, Multi-Layer Perceptron, and Voting Classifier, used to determine the reliable and accurate factors common with Alzheimer's disease. Feature selection and feature scaling are applied in this work when selecting the best attributes to enhance the models' accuracy. The proposed study provides a more beneficial outcome; FFNN achieved the highest accuracy of 87.84% with a 60% training split and maintained strong performance across other splits. CNN also showed excellent results, reaching up to 89.38% accuracy with a 75% training split and maintaining competitive performance with other splits. SVM demonstrated strong performance with an accuracy of 93.40% on an 80% training split, making it the best-performing machine learning model. RF also performed well, achieving up to 88.49% accuracy with a 75% training split. The Soft Voting model has the highest accuracy of 94.50%, showing excellent discriminative ability.

Future studies will pay particular attention to attempting to delete unnecessary features from existing feature sets and utilize extractions and analysis of potential unique features that would help in the detection of Alzheimer's disease, use diverse datasets, and perform multiclass classification at each level of AD.

6. References

1. Z. Breijyeh and R. Karaman, "Comprehensive Review on Alzheimer's Disease: Causes and Treatment," Molecules, vol. 25, no. 24, pp. 5789, Dec. 2020.
2. WHO, "Dementia key facts," 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia. [Accessed: Jan 20, 2024].
3. Physicians PC, "2020 Alzheimer's disease facts and figures," Alzheimer's Dement., vol. 16, no. 3, pp. 391–460, 2020.
4. K. M. M. Uddin, M. J. Alam, J.-E. Anawar, M. A. Uddin, and S. Aryal, "A novel approach utilizing machine learning for the early diagnosis of Alzheimer's disease," *Biomedical Materials & Devices*, Feb. 2, 2023.
5. J. Tcw and A.M. Goate, "Genetics of Beta-Amyloid Precursor Protein in Alzheimer's Disease," Cold Spring Harb. Perspect. Med., vol. 7, no. 6, art. no. a024539, 2017.
6. C. Dailey, "The Impact of Alzheimer's Disease - The Silent Killer," JCCC Honors Journal, vol. 7, no. 2, 2016.
7. G. W. Small, "Neuroimaging and genetic assessment for early diagnosis of Alzheimer's disease," *The Journal of Clinical Psychiatry*, vol. 57, pp. 9-13, 1996.
8. University of Utah Health Sciences Center, "PET Scan Distinguishes Alzheimer's from Other Dementia," *ScienceDaily*, Nov. 1, 2007. [Online]. Available: https://www.sciencedaily.com/releases/2007/11/071101112655.htm. [Accessed: Jan. 11, 2024].
9. L. Beyenhof, "How a CT Scan Works," Helium - Where Knowledge Rules. [Online]. Available: http://www.helium.com/items/302203-how-an-ct-scan-works. [Accessed: 02-Jan-2024].
10. K. M. Gosche *et al.*, "Hippocampal volume as an index of Alzheimer neuropathology: Findings from the Nun Study," *Neurology*, vol. 58, pp. 1476-1482, 2002.
11. University of South Florida Health, "MRI Brain Scans Accurate In Early Diagnosis Of Alzheimer's Disease," *ScienceDaily*, 23 Dec. 2008. [Online]. Available: http://www.sciencedaily.com/releases/2008/12/081218213630.htm. [Accessed: 11-Jan-2024].

Online ISSN 3007-3197
Print ISSN 3007-3189

12. "Alzheimer's Disease," Buoy Health, [Online]. Available: https://www.buoyhealth.com/learn/alzheimers-disease#definition. [Accessed: 07-Jan-2024].
13. S. Afzal, M. Maqsood, U. Khan, and I. Mehmood, "Alzheimer Disease Detection Techniques and Methods: A Review," International Journal of Interactive Multimedia and Artificial Intelligence, vol. (In Press), pp. 1, 2021
14. Burns, "Diagnosis and Management of Alzheimer's Disease," Dialogues Clin. Neurosci., vol. 2, no. 2, pp. 129–138, Jun. 2000.
15. K. Mahapatra, "Detection of Alzheimer's Disease Using MRI Scans Based on Inertia Tensor and Machine Learning," arXiv preprint arXiv:2304.13314, 2023.
16. D. Agarwal, M. Á. Berbís, A. Luna, V. Lipari, J. B. Ballester, and I. de la Torre-Díez, "Automated Medical Diagnosis of Alzheimer's Disease Using an Efficient Net Convolutional Neural Network," Journal of Medical Systems, vol. 47, no. 1, pp. 10916-023-01941-4, 2023.
17. R. Prajapati, U. Khatri, and G.R. Kwon, "An Efficient Deep Neural Network Binary Classifier for Alzheimer's Disease Classification," International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 231–234, IEEE, 2021.
18. S. Basheer, S. Bhatia, S.B. Sakri, "Computational Modeling of Dementia Prediction Using Deep Neural Network: Analysis on OASIS Dataset," IEEE Access, vol. 9, pp. 42449–42462, 2021.
19. S. Spasov, et al., "A parameter-efficient deep learning approach to predict conversion
20. from mild cognitive impairment to Alzheimer's disease," Neuroimage, vol. 189, pp. 276–287, 2019.
21. K. Aderghal, et al., "Classification of Alzheimer's disease on imaging modalities with deep CNNs using cross-modal transfer learning," in 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), Piscataway, 2018, pp. 345–350.