



A Novel Approach Utilizing Machine Learning for the Early Diagnosis of Alzheimer's Disease

Khandaker Mohammad Mohi Uddin¹ · Mir Jafikul Alam¹ · Jannat-E-Anawar¹ · Md Ashraf Uddin² · Sunil Aryal²

Received: 2 February 2023 / Accepted: 27 March 2023 / Published online: 10 April 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC 2023

Abstract

Alzheimer's disease (AD) is one of the leading causes of dementia among older people. In addition, a considerable portion of the world's population suffers from metabolic problems, such as Alzheimer's disease and diabetes. Alzheimer's disease affects the brain in a degenerative manner. As the elderly population grows, this illness can cause more people to become inactive by impairing their memory and physical functionality. This might impact their family members and the financial, economic, and social spheres. Researchers have recently investigated different machine learning and deep learning approaches to detect such diseases at an earlier stage. Early diagnosis and treatment of AD help patients to recover from it successfully and with the least harm. This paper proposes a machine learning model that comprises GaussianNB, Decision Tree, Random Forest, XGBoost, Voting Classifier, and GradientBoost to predict Alzheimer's disease. The model is trained using the open access series of imaging studies (OASIS) dataset to evaluate the performance in terms of accuracy, precision, recall, and F1 score. Our findings showed that the voting classifier attained the highest validation accuracy of 96% for the AD dataset. Therefore, ML algorithms have the potential to drastically lower Alzheimer's disease annual mortality rates through accurate detection.

Keywords Healthcare · Prediction · Alzheimer's disease (AD) · Machine learning · Feature selection · OASIS

Introduction

Alzheimer's disease (AD) [1] is the most prevalent neurological disease that causes dementia in older adults over the age of 65, affecting around 6.5 million people worldwide [2]. AD is a subtype of dementia rather than its cause. Dementia is an all-encompassing word referring to a decrease in cognitive function, including memory, thinking, and reasoning,

which is severe enough to interfere with daily life. AD is the leading cause of dementia, accounting for between 60 and 80% of cases. Memory loss, difficulties with language and communication, disorientation, mood changes, and finally difficulty with basic self-care chores such as dressing and bathing are typical signs of Alzheimer's disease. It is considered that a mix of genetic, environmental, and behavioral factors contribute to the development of AD. There is currently no cure for AD, although there are therapies that can help control symptoms and decrease the illness's progression. These treatments may include drugs to improve cognitive function, therapy to address behavioral and psychological issues, and improvements to one's lifestyle, such as physical activity and a balanced diet [3].

Alzheimer's disease typically begins with mild memory problems that progressively worsen over time, leading to impaired brain function. While the exact cause of Alzheimer's is not fully understood, there are several factors that are thought to contribute to its development, including aging, genetic predisposition, untreated clinical depression, lifestyle factors, severe head injury, and prolonged hypertension.

✉ Khandaker Mohammad Mohi Uddin
jilanicsejnu@gmail.com

Mir Jafikul Alam
jafialam@gmail.com

Jannat-E-Anawar
jannateanawar@gmail.com

Md Ashraf Uddin
ashraf.uddin@deakin.edu.au

Sunil Aryal
sunil.aryal@deakin.edu.au

¹ Department of Computer Science and Engineering, Dhaka International University, Dhaka 1205, Bangladesh

² School of Information and Technology, Deakin University, Warun Ponds, Geelong, Australia

The human brain is composed of billions of neurons that form connections with each other. In Alzheimer's disease, these connections are lost due to the buildup of abnormal protein structures known as “plaques” and “tangles,” which ultimately led to the death of neurons. Plaques are deposits of amyloid beta (Abeta), [4] a peptide that is insoluble. Alzheimer's disease is typically divided into three phases: early, middle, and late. In the late stage, individuals may develop dementia.

In the early stage of Alzheimer's disease, a person may experience repeated forgetfulness, which differs from normal forgetfulness. While it is common for anyone to forget things, in the early stage of Alzheimer's disease, a person's brain capacity progressively declines, making it difficult to remember regular activities. Confusion is also common in the early stage, whereby a person may forget what they were doing or what they intended to do. Some of the common challenges in the early stage of Alzheimer's disease include [5]

- Difficulty recalling recent conversations or events.
- Frequently misplacing items.
- Struggling to remember the names of places and things.
- Having trouble finding the right words.
- Making poor judgments or struggling to make decisions.
- Becoming less adaptable and more resistant to change.
- Experiencing memory issues that interfere with daily activities.
- Finding it difficult to solve problems or plan ahead.
- Having trouble completing routine tasks.

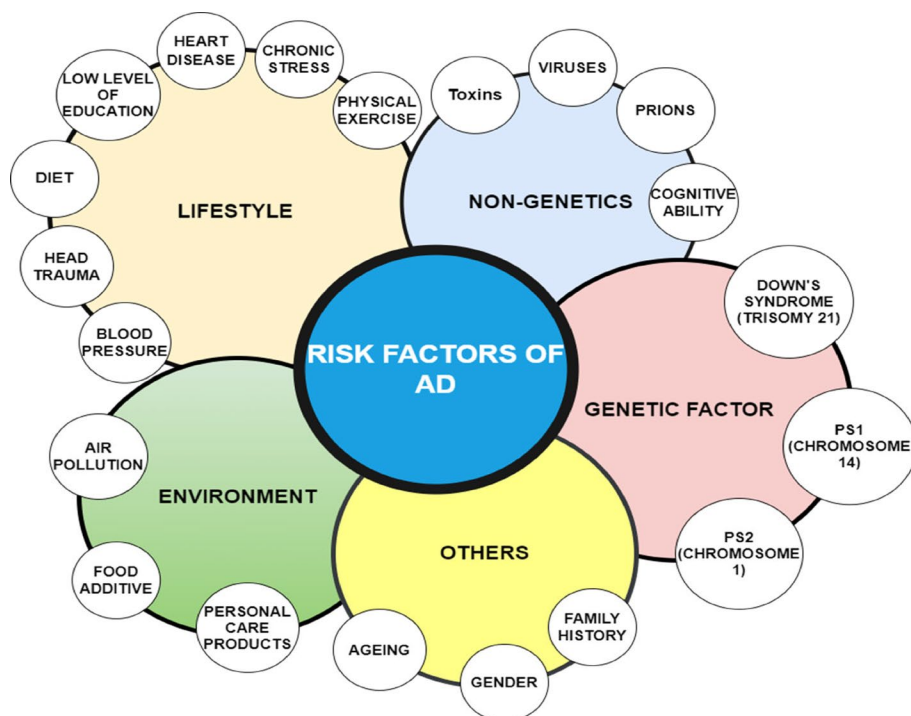
- Confusion about time or location.
- Losing track of items and the ability to recall past events.

During the middle stage of Alzheimer's disease, confusion tends to worsen gradually, and patients may also experience difficulties with sleep [6]. In the late stage, individuals may become less responsive to their environment, struggle with communication, and lose control over their movements. Although they may still be able to express words or phrases, they may have trouble conveying their emotions. The risk factors of AD are depicted in Fig. 1.

Recently researchers and academician have shown interest in computer-assisted machine learning methodologies for analyzing and predicting different diseases using medical data. Many state-of-the-art works applied traditional pattern analysis techniques, such as linear discriminant analysis (LDA), linear program boosting method (LPBM), logistic regression (LR), support vector machine (SVM), and support vector machine-recursive feature elimination (SVM-RFE). The outcomes from these approaches are promising for the early detection of Alzheimer's disease and the prediction of AD progression [7].

To implement machine learning models, we require designing architecture and perform preprocessing. In general, machine learning-based classification studies involve four steps: feature extraction, feature selection, dimensionality reduction, and feature-based classification algorithms. These processes require specialized knowledge and several optimization steps, which can be time consuming. The reproducibility of these methods has been a problem [8]. In

Fig. 1 Risk factor of Alzheimer's disease (AD)



the feature selection procedure, for instance, AD-related features are selected from various neuroimaging modalities to derive more informative combinatorial measures, which may include mean subcortical volumes, gray matter densities, cortical thickness, brain glucose metabolism, and cerebral amyloid accumulation in regions of interest (ROIs) such as the hippocampus [9]. The fast development in high-volume biomedical datasets (neuroimaging and related biological data) over the past decade, in tandem with the advancements in machine learning (ML), has opened up new pathways for the diagnosis and prognosis of neurodegenerative and neuropsychiatric illnesses [10, 11]. In this paper, we adopt an ensemble-based machine learning approach to detect AD.

The rest of the paper is organized as follows. In "Review of Machine Learning Usage in Medical Clinical Diagnostics" section, we review the literature. The methodology is presented in "Materials and Methods" section. The result is analyzed with respect to different performance metrics in "Results and Discussions" section before concluding the paper in "Conclusions" section.

Review of Machine Learning Usage in Medical Clinical Diagnostics

In the field of medical science, a number of applications that use various machine learning techniques are presently being used for data analysis and innovation. Machine learning techniques have been used in a number of recent healthcare research studies, including the diagnosis of COVID-19 using X-rays [12, 13], the identification of tumors using MRIs [14, 15], the prediction of cardiovascular diseases [16, 17], dengue [18, 19], stroke [20], and cancer [21, 22]. Kader et al. [23] created a model that employed feature selection and extraction strategy to predict Alzheimer's disease using machine learning techniques. The OASIS longitudinal dataset is used for classification. A concise introduction of the many methods used to analyze brain scans in order to identify brain illnesses. This research also showed the technique that is most reliable for identifying brain diseases. Using 22 brain disease databases, the authors can identify the most precise diagnostic technique. The research combines current findings on Alzheimer's disease, Parkinson's disease, epilepsy, and brain malignancies.

Mehmood et al. [24] presented an overview of recent studies on the categorization and Alzheimer's disease diagnosis. It offers instances of the methodology used to identify and classify AD. Martinez-Murcia et al. [25] developed a model in which deep convolutional autoencoders are used to examine data analysis of AD. This model can extract MRI features from MRI images. It characterized a person's cognitive problems and the underlying neurodegenerative process via data-driven deconstruction. Imaging-derived

markers and MMSE or ADAS11 scores can be used to predict an AD diagnosis in more than 80% of cases. Helaly et al. [26] developed a model which employed coherent approach for Alzheimer's disease early detection. This study utilized convolutional neural networks (CNN) to categorize AD. Two techniques are applied to predict AD. Using a web application, clinicians and patients may remotely screen for Alzheimer's disease. According to the AD spectrum, it also establishes the patient's AD stage. The VGG19 pre-trained model has been improved, and it now identifies AD stages with an accuracy of 97%.

Kavitha et al. [27] constructed a model which uses optimal parameters for predicting Alzheimer's disease. Classifiers like Gradient Boosting, Support Vector Machine, and Decision Tree and Voting are used to identify AD. The proposed study displays excellent results, with 83% average validation accuracy. This test accuracy score is significantly higher than that of prior efforts.

Ghazal et al. [28] developed a model which utilized transfer learning on multiclass categorization using brain MRIs, to classifying the pictures into four categories: very mild dementia (VMD), mild dementia (MD), and moderate dementia (MOD) non-dementia (ND). The correctness of the proposed system model is 91.70% according to simulation findings. It was also noted that the suggested technique provides findings that are more accurate when compared to earlier methods.

Gaudiuso et al. [29] developed a model which utilizes Laser-Induced Breakdown Spectroscopy (LIBS) and machine learning. This study examined micro-drops of plasma samples from AD and healthy controls (HC) results in robust categorization. After obtaining the LIBS spectra of 67 plasma samples from a population of 31 AD patients and 36 healthy controls (HC), we successfully detected late-onset AD (> 65 years old), with a total classification accuracy of 80%, specificity of 75%, and sensitivity of 85%.

Nawaz et al. [30] developed a model which utilized deep feature-based strategy for predicting the stage of Alzheimer's disease using a convolutional neural network. By transferring the initial layers of a previously trained Alex Net model and extracting the deep features Researchers applied the popular Machine learning approaches include k-nearest neighbor (KNN), random forest, and support vector machine (SVM) for the categorization of the retrieved deep features (RF). According to the assessment findings of the suggested plan, a deep feature-based model beat hand-crafted and deep learning techniques with an accuracy rate of 99.21%. Basheer et al. [31] constructed a model using the OASIS dataset. Demented and non-demented classifications were used in the study to distinguish the classes. By proving the correlation accuracy across a number of repetitions with an acceptable accuracy of 92.39%, the discovery has been confirmed as accurate.

Prajapati et al. [32] developed a model using deep neural networks. The study has produced improved categorization outcomes in contemporary medical research sectors. The suggested DNN with the highest validation accuracy score achieved test data accuracy of 85.19%, 76.93%, and 72.73%. Lucas et al. [33] developed model based on quantitative EEG (qEEG) processing technique to automatically distinguish AD patients from healthy people. 19 healthy patients and 16 AD patients with probable mild to moderate symptoms had their EEGs analyzed. The accuracy and sensitivity of the analysis, which took into account each patient's particular diagnosis, were 87.0% and 91.7%, respectively. The accuracy and sensitivity of the EEG analysis were found to be 79.9% and 83.2% respectively. The accuracy and sensitivity of the analysis, which took into account each patient's particular diagnosis, were 87.0% and 91.7%, respectively. Sudharsan et al. [34] created a model using structural magnetic resonance imaging (sMR). The study utilize moderate cognitive impairment (MCI), and healthy control patients, to diagnosis early Alzheimer's disease.

The main contributions to this proposed research work are highlighted as follows:

- To employ a feature selection approach to identify the most relevant features and avoid data redundancy. The feature selection method chooses the most important k number of features, and we apply a standard scaler for scaling features' values.
- To train the model using the oasis dataset that contains numerous missing values. We utilized the mean approach to address them.
- To predict AD, we apply six different machine learning classifiers and propose a voting-based ensemble approach which demonstrates around 96% accuracy.

Materials and Methods

The three basic steps of the proposed technique are data collection, preprocessing, and prediction utilizing machine learning algorithms to forecast Alzheimer's disease. Pandas [35] is used to load the initial dataset and import the required libraries for preprocessing. The consistency and redundancy of the input dataset reduce the machine learning algorithm's accuracy. For optimal outcomes, in this research, the data are cleaned up before being used in a machine learning algorithm, eliminating unnecessary values and attributes. The data have been preprocessed and randomly spilled into testing and training. The ratio of dividing data is 80:20 randomly which means 80% of the data are used to train the model and 20% for testing. Figure 2 illustrates the system's process for making early prediction of Alzheimer's disease.

Dataset

A longitudinal dataset [36] is utilized in this research to predict Alzheimer's disease. The initial task is to ascertain how cross-sectional the data at a given period or at a certain baseline. Following that, a thorough data analysis is carried out, which included comparing the primary research components and the associated data obtained on each visit. The research includes 150 individuals with MRI data ranging in age from 60 to 96. Each patient was scanned at least once during the study. Every patient is right handed. At the time of the preliminary examination, 64 patients are identified with Dementia, whereas 72 are classed as non-demented, which stayed the same during the research.

Table 1 shows the OASIS longitudinal dataset description. The attribute visits indicate the number of patients visited at any point throughout the trial. M/F specify the patients' gender M and F stand for male and female, respectively. The age of the patients is described via the attribute age. Patients' study time is determined by the EDUC characteristic. SES stands for patients' socioeconomic status. Mini-Mental State Examination and nWBV are the two terms normalize the entire brain's volume.

Data Preprocessing

The raw dataset has data redundancy and missing value [37, 38]. As part of data handling, missing value features are extracted and transformed. Feature selection and feature scaling are also performed in this work for the preprocessing of the dataset.

Data Analysis

Data analysis is inspecting the dataset, then cleansing, transforming, and modeling the dataset into a suitable form. Data analysis [39] aims to find helpful information that will later be used to support decision making. For example, the analysis could help in evaluating the characteristics of the data and the relation between co-relation attributes. Table 2 displays the minimum, maximum, and median values for the each attribute of the dataset.

Missing Data Handling

The OASIS dataset contains a number of missing values. Missing values can affect the outcomes ML model or reduce the model accuracy. In this study, we employed the mean method to impute missing values. We obtained the mean (average) of all observable data and replaced it with the mean values in place of missing values.

Fig. 2 Working flowchat of the proposed system to predict AD

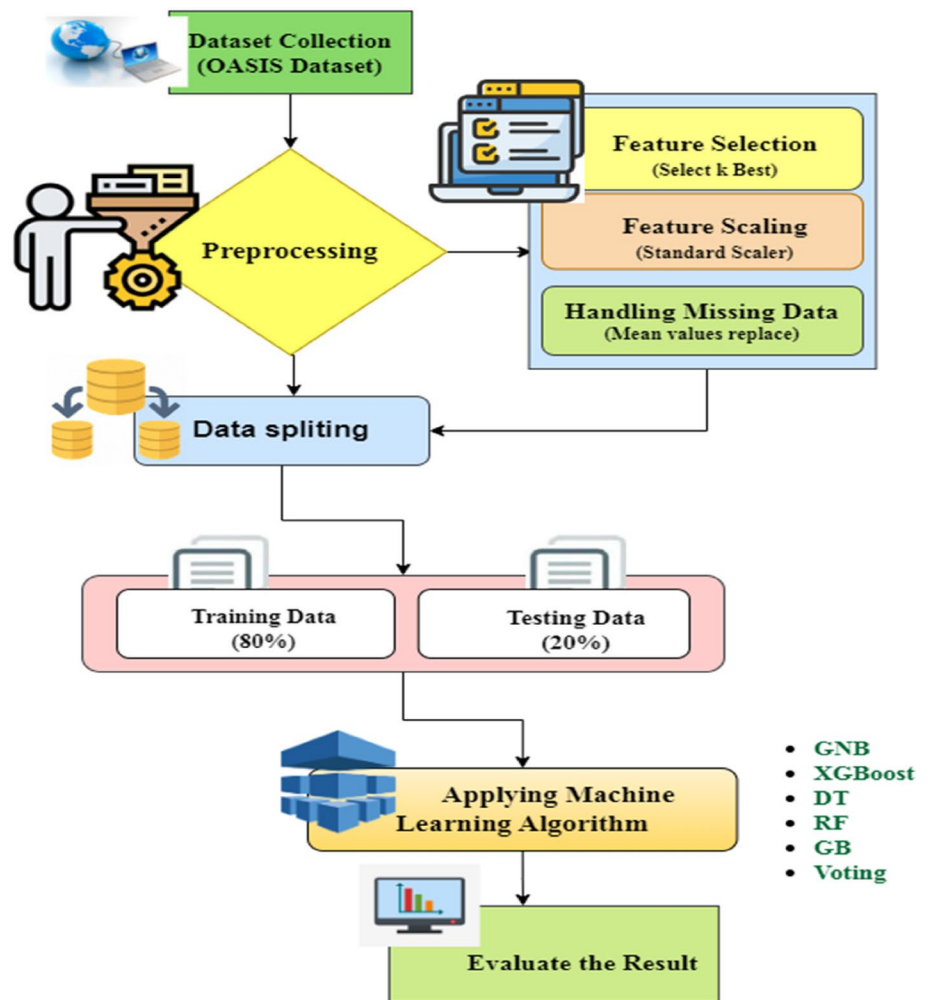


Table 1 Description of the dataset

Sl. No	Attribute	Description
1	Visit	Number of visits during study
2	MR delay	Delay
3	M/F	Sex (F=Female, M=Male)
4	Age	Age in (years)
5	EDUC	Education in (years)
6	SES	Social and economic status
7	MMSE	Examination of mini-mental state
8	CDR	Clinical dementia rating
9	eTIV	Estimated total incremental value
10	nWBV	Normalize total brain volume
11	ASF	Atlas scaling factor

Table 2 Each attribute's maximum, minimum, and median values

Attribute	Min	Max	Mean	Median
EDUC	6	23	14.59	15
MMSE	4	30	27.35	29
SES	1	5	2.43	2.0
CDR	0	2	0.29	0
nWBV	0.64	0.83	0.72	0.729
ASF	0.87	1.58	1.19	1.194
eTIV	1106.0	2004	1488	1470

$$\text{Mean} = \frac{\text{Sum of all data points}}{\text{Number of data points}} \tag{1}$$

In column “SES,” 19 rows of missing values have been found, and “MMSE” is two rows of missing values found.

Imputation is a method of replacing missing values by replacing them with equivalent values. The 19 rows of missing value have been replaced with mean value. As for “MMSE,” 2 rows of missing values have been replaced with the Mode Value. Figure 3 displays the data visualization before and missing data handling.

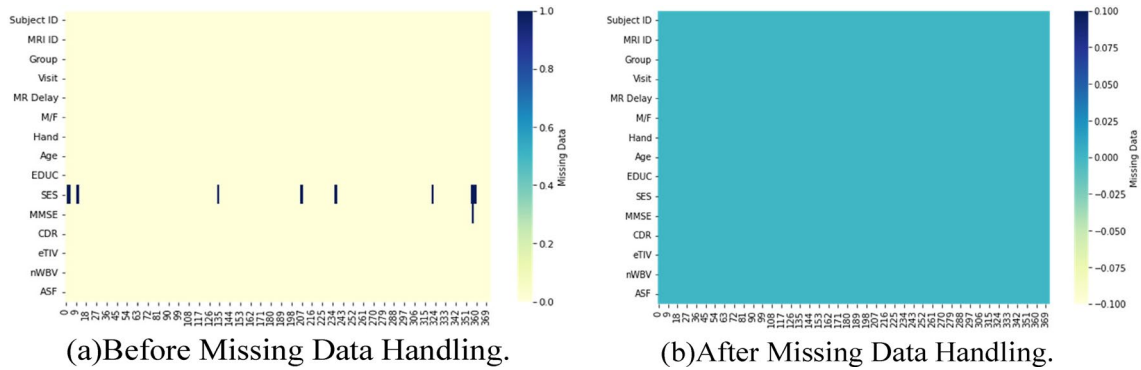


Fig. 3 Display the data visualization. **a** Before missing data handling, **b** after missing data handling

Table 3 Sore value of each attribute using “score func”

Sl.No	Attribute	Score value
1	Visit	3.234565
2	MR delay	6.627021
3	M/F	13.413107
4	Age	3.132662
5	EDUC	13.485719
6	SES	11.530848
7	MMSE	107.216426
8	CDR	436.908989
9	eTIV	0.674019
10	nWBV	20.581034
11	ASF	0.381442
12	Subject ID	0.058990
13	MRI ID	0.054917

Feature Selection

While creating a predictive model, feature selection is the process of choosing relevant features by reducing the dimensionality of the input attribute. Feature selection [40] is crucial in machine learning. The purpose of feature selection techniques is to decrease the quantity of input parameters. Therefore, it is beneficial to a model to predict and increase the accuracy score by reducing identify of the dataset. In this research, feature selection is utilized to analyze clinical data related to Alzheimer's disease. Select kbest choose features are based on the k highest scores. In this work, the select kbest approach is used to identify the best feature, and f class is used as the scoring function. This method is employed since it allows us to use both classification and regression data by just changing the “score func” option. In Eq. (2), x represents the raw score, μ represents the population's mean, and σ denotes the standard deviation. Table 3 represents the score value of each attribute using “score func.”

$$Z = \frac{x - \mu}{\sigma} \tag{2}$$

Correlation Coefficient

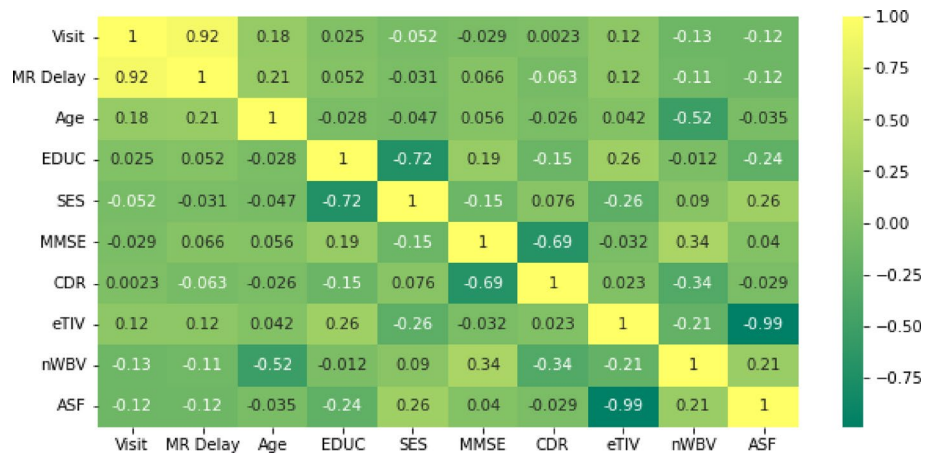
The linear connection between two variables is measured by their covariance. It is simple to discover a link between the different phases of Alzheimer's using correlation coefficients. The drawback with this strategy is that the information is gathered from a variety of sources, making it subject to outliers. Equation (3) is used to define the correlation between the two variables J and M [41].

$$\rho_{J,M} = \frac{\text{cov}(J, M)}{\sigma_J \sigma_M} \tag{3}$$

Heat-Map

A heat map is a type of data visualization method showing a phenomenon’s magnitude in two dimensions using color. The reader can better see how the phenomena are grouped or fluctuate in space thanks to the color variation [42]. The values of the first dimension are presented as rows in the table, while the values of the second dimension are displayed as columns. The color of the cell is determined by the proportion of measurements that match the dimensional value. The pattern emphasizes distinctions and variances within the same dataset. In Fig. 4, warmer color indicates higher values and colder color indicates lower values. Row attribute visit is correlated with column visit which is shown in warmer color yellow and values is 1 which is highest MR Delay and visit is highly correlated that is shown in warm yellow color. ASF and visit attribute has lower correlation which is shown colder color light green. In Fig. 4, same attribute is highly correlated whereas Visit and ASF nWBV and Visit eTiv and Visit have lower correlation.

Fig. 4 Heat-map for correlation between every attribute



Standard Scaler

Data standardization is the process of combining several datasets' structures into a single uniform data format. Standardizing a dataset includes rescaling data that mean observed values is 0 and standard deviation is 1. Standard scaler is used in this work.

Classifier Models

Gaussian Naive Bayes (Gaussian NB)

Gaussian Naive Bayes is a simple probabilistic algorithm. This is one of the most widely used Naive Bayes (NB) algorithm that utilizes the Bayes theorem. The strategy is made to deal with the continual qualities that are to associate each class and generated based on a Gaussian distribution. The NB family has several key benefits, including the ability to be trained extremely successfully in supervised learning, the ability to be applied to real-world classification problems, and the need for minimal training data. The NB family has a significant flaw in that the qualities are expected to be independent, which is virtually impossible. Thus to calculate the probability of a continuous data set, the following Eq. (4) can be used [43].

$$P(X = x|C = c) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4}$$

where x = variable, c = class, π = mean, σ = stander deviation.

XGBoost

XGBoost stands for Extreme Gradient Boosting. It works at its fastest and most efficient utilizing gradient-boosted decision trees. Here, D represent the dataset and M represent the number of training samples and i ranges from 1

to m where $\{x_i, y_i\}$ represents the i th training example. In Eq. (5) the estimated label is performed.

$$\hat{y}_i = \vartheta(x_i) = \sum_{M=1}^M f_M(x_i), f_M \in P \tag{5}$$

where P is the space of decision trees which is also known as Classification and Regression Tree (CART). Each f_M corresponds to an independent tree structure. For boosting tree algorithm, Eq. (6) is the regularized objective function minimized in which $\Omega(f)$ represents the $L1$ regularization. Besides, l is a differentiable convex loss function.

$$(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_M \Omega(f_M). \tag{6}$$

For gradient boosting algorithms, Eq. (7) is the objective function. $y_i^{(t)}$ is the estimation of the i th instance at the t th iteration.

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t). \tag{7}$$

Equation (8) shows the second-order approximation which is used to optimize the objective Falster.

$$\bar{\mathcal{L}}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t). \tag{8}$$

In Eq. (7), $g_i = \delta_y^{(t-1)} l(y_i, \hat{y}_i^{(t-1)})$ and $h_i = \partial_{y^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)})$ are the 1st and 2nd-order statistics on the lost function [44].

Decision Tree (DT)

The goal of DT is to predict the value of a target variable. A DT is utilized, with the leaf node representing a class label and the interior node representing features [45]. Equations (10, 11) help for calculating the output of decision tree.

Information gain

$$\text{Gain}(M, J) = \text{Entropy}(M) - \sum_{v \in \text{Values}(J)} (|M_v|/|M|) \text{Entropy}(M_v) \tag{9}$$

Gain Ratio: Gain Ratio $(M, J) = \text{Gain}(M, J) / \text{Split Information}(M, J)$

$$\text{SplitInformation}(M, J) = \sum_{i=1}^c (|M_i|/|M|) \log_2(|M_i|/|M|). \tag{10}$$

Gini value

$$\text{Gini}(D) = 1 - \sum_{j=1}^n p_j^2. \tag{11}$$

where p_j is relative frequency of class j in D . If dataset D is split on M into two subsets $D1, D2$ the gini index $\text{gini}(D)$ is defined as follows: $\text{GiniA}(D) = |D1|/(|D| \text{gini}(D1)) + |D2|/(|D| \text{gini}(D2))$.

Random Forest (RF)

A random forest model outperforms a decision tree model because it avoids the problem of overfitting. Random forest models are composed of a variety of decision trees, each completely distinct from the others. In Random Forest, several trees are utilized to build a forest or forest, and each tree is then continually evaluated [46]. Equation 12 is used to get the Gini Index for the classification:

$$\text{Gini} = 1 - \sum_{i=1}^n (p_i)^2. \tag{12}$$

In Eq. (12), the value of p_i is the probability of the object to be classified in a certain class/feature.

Gradient Boosting

Both the base-learner models and the loss function are freely securable. When given a particular loss function (y, f) and/or a particular base-learner (x, θ_t) , the answer to the parameter estimates could be complicated to calculate in actuality [47]. It was suggested to address this by choosing a different function $h(x, \theta_t)$ that is most parallel to the observed data's negative gradient, $g_t(x) \text{Ni} = 1$.

$$g_t(x) = \text{Ey} \left[\frac{\partial \psi(y, f(x))}{\partial f(x)} \Big|_x \right] f(x) = f^{\wedge}t - 1(x). \tag{13}$$

Voting

One of the simplest methods of merging the forecasts from several learning algorithms is by voting. Voting classifiers are not really classifiers, but rather wrappers for multiple ones that are trained and evaluated concurrently to benefit from their unique qualities. To predict the final result, datasets are trained using various algorithms and ensembles. A qualified majority on a forecast can be obtained in two ways:

Hard Voting Hard voting is the simplest type of majority voting. The class with the most votes (N_c) will be selected in this instance. The majority vote of each classifier is used to make prediction [48]. In Eq. (14), the class label j is predicted using majority (plurality) voting of each classifier M

$$j = \text{mode} \{M_1(x), M_2(x), \dots, M_n(x)\}. \tag{14}$$

Model Validation

In this study, overfitting issue is diminished via model validation. To measure the model accuracy, Cross-Validation is used to train the ML models. Moreover, making the ML model noise-free is a daunting task. Hence, the proposed study uses 10-fold cross-validation method, which divides the whole dataset into 10 divisions those are equal in size. The ML model trains each iteration using the 9 divisions. The performance of the approach is examined using the mean of all 10-folds.

Results and Discussions

In this work, different performance metrics are used including F1 score, recall, accuracy, and precision. 10-fold cross-validation approach is utilized to get each model's ideal parameter. After that, each model's accuracy is evaluated. The confusion matrix is used to explain performance assessments, which can be either binary or multiclass in nature. A unique machine learning classifier is constructed and validated to predict and differentiate actual Alzheimer's disease-affected persons, and a learning model is created to distinguish truly afflicted individuals from a given population. These features are used to compute the precision, recall, accuracy, and F-score assessment metrics. According to this study, the recall (sensitivity) is the percentage of individuals who are correctly categorized as having Alzheimer's. The percentage of patients who are accurately identified as

not having the condition indicates the accuracy of an Alzheimer's diagnosis. In contrast, accuracy denotes the percentage of subjects that are properly identified, whereas F1 denotes the weighted average of recall and precision. The patient is given a report that details the findings and the stage of Alzheimer's disease they are now experiencing. Because the phases are dependent on the patients' responses, it is crucial to identify them. In addition, recognizing the stage helps doctors in recognizing how the disease is impacting patients. For the purpose of executing its tests and data analysis, this study applied the following settings, resources, and libraries:

- (a) Environments Used:- Python 3
- (b) Scikit-learn libraries for machine learning

Figure 5 shows that Men are more likely than women to have dementia or Alzheimer's disease. Figure 6 shows that compared to the dementia group, the non-demented group

had significantly better MMSE (Mini-Mental State Examination) scores.

Figure 7a–c indicates resolution values of ASF, eTIV, and nWBV for Demented and Non-demented group of people. The ratio of brain volume that the non-demented group is larger than the demented group, as shown by the graph in Fig. 7. Figure 8 indicates the analyzed results which denoted for demented and Non-demented people of EDUC.

Figure 8 shows the analysis of education on years and Fig. 9 illustrates the age factor to determine the proportion of afflicted individuals based on the demented and non-demented groups. It has been shown that people with dementia tend to be older on average between 70 and 80 years old than patients without dementia. People who suffer from such type of disease are probably not very likely to survive. There are not many people that are above 90 years old.

The following is a summary of the intermediate findings from the analysis of the attributes shown above.

Fig. 5 Resolution of demented and non-demented rate based on gender, Gender group Female=0, Male=1 and Non Dementia=2

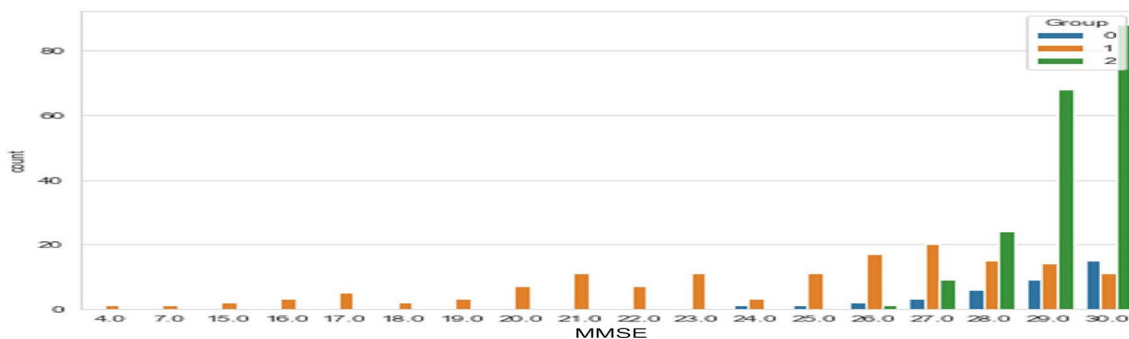
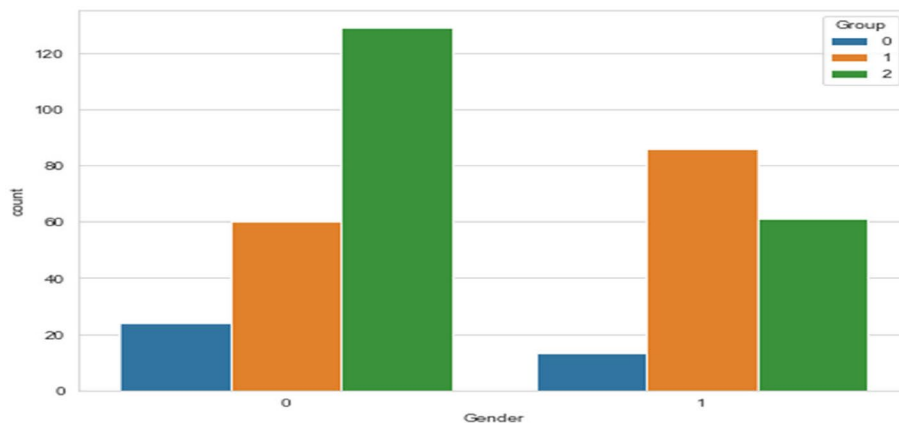


Fig. 6 Resolution of MMSE scores for demented (Female and Male) and non-demented group of patients

Fig. 7 a–c Resolution of ASF, eTIV and nWBV for Demented and Non-demented group

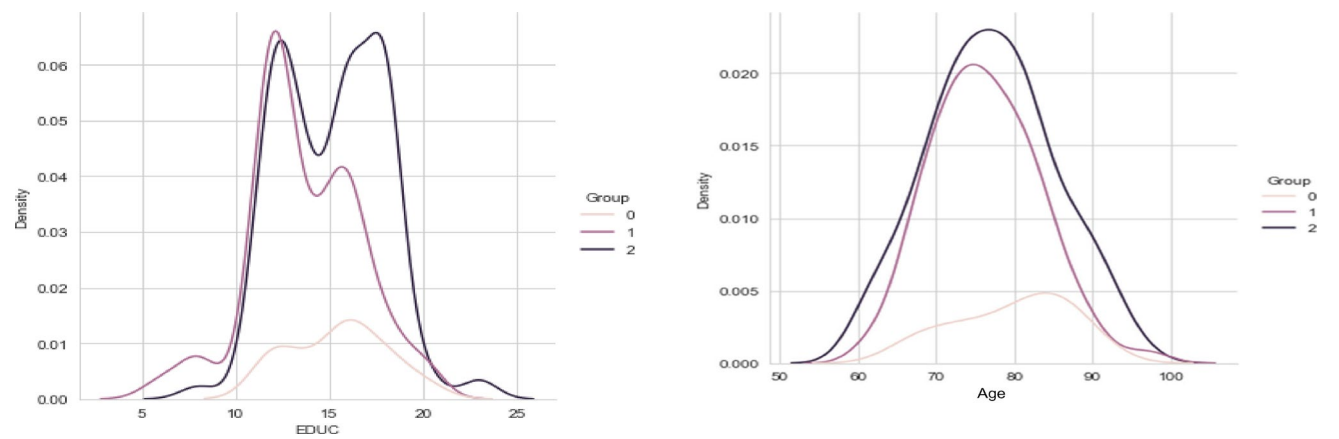
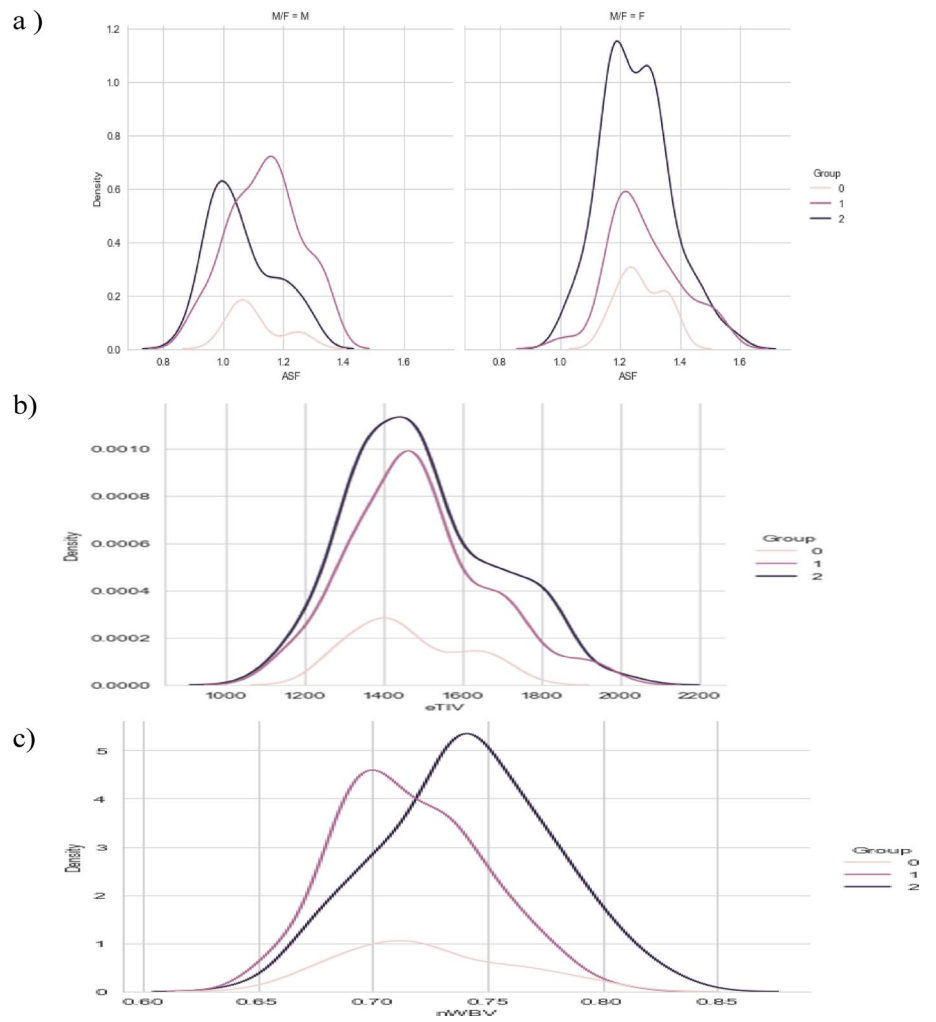


Fig. 8 Resolution of education on years

Fig. 9 Resolution on people impacted by demented and non-demented group based on age

1. Men are more likely than women to have dementia or Alzheimer's disease.
2. Patients with dementia have less education in regards to educational years.
3. There is a difference in brain volume between those with and without dementia.
4. The demented group had a higher number of patients in their 70 s and 80 s as compared to the non-demented group.

Performance Evaluation Measures

Here, the following performance assessment metrics (Eqs. 15–18) are computed using the true-positive (TP), false-positive (FP), true-negative (TN), and false-negative (FN) counts.

Accuracy Finding the percentage of correctly categorized outcomes from all examples produces this measurement.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (15)$$

Precision This method evaluates the proportion of accurately predicted positive rates to all predicted positive rates. When the precision value is 1, the classifier is considered to be effective.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (16)$$

Recall Recall can define as a true-positive rate. Whether the recall is 1, it is significant as a good classifier.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (17)$$

F1 Score It is a measurement that takes the Recall and Precision parameters into account. Only when both measures, such as recall and precision, are 1, does the F1 score become 1.

$$F1 = 2 \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (18)$$

Precision and recall are the primary performance indicators utilized in medical diagnostics research. Because it combines recall and accuracy into a single number that is easier to use for comparisons, the F-measure is especially significant for reporting performance in medical diagnostics. Recall is viewed as a more significant parameter in the medical field since a false negative is typically more damaging

than a false positive. A missing condition might have severe repercussions for a patient, but a missed FP (precision) might not be as essential as a missed FN (recall) because FPs can be dismissed by doctors. Given this, a modified version of F1 that emphasizes recall over precision is a more useful statistic. The confusion matrix ML models are shown in Fig. 10.

Table 4 shows that training and testing accuracies for each model have been compared to minimize overfitting. Precision, recall, accuracy, and F1 score for each model are also presented in Table 4. The results of the study shown in Table 03 affirmed that the best and optimal procedures, which have stellar results, are random forest, and GaussianNB Voting classifier.

Figure 11 shows the graphical comparison of evaluation metrics attained by the proposed system and existing classifiers: for GaussianNB, Decision Tree, Random Forest, XGBoost, GradientBoost, and Voting Classifier. Here shows the comparison among accuracy, precision, recall, and F1 Score using ML algorithms where easily noted the highest value and lowest value.

Figure 12 denotes ROC curve for multiclass classifiers where blue color defines as micro-average ROC curve, and yellow, green, and red define as ROC curve of class for multiple class (class-0, class-01, class-2) for (a) GaussianNB (b) Decision Tree (c) Random Forest (d) XGBoost (e) GradientBoost (f) Voting Classifier.

Figure 13 demonstrates that the result of test dataset on AD where the characteristics with the lowest squared error identified by the model. All the missing values are removed through SGD learning, which also changed the nominal properties into binary ones. Also, every attribute is normalized MLP learning.

Comparison with Existing Works

Table 5 shows Comparison with the existing work for AD prediction using ML Martinez-Murcia et al. [25], and Basheer et al. [31] used both machine learning and deep learning. Sudharsan et al. [34] used machine learning technique and PCA and got 78.31% of accuracy to predict AD. Kavitha et al. [27] used OASIS dataset to predict AD, and using ML, they obtained 83% accuracy. OASIS dataset is also used by Basheer et al. [31] to predict AD with machine learning and deep learning models. The authors obtained the accuracy of 92.39%. As far as it can observe that the suggested method outperforms all others in the literature.

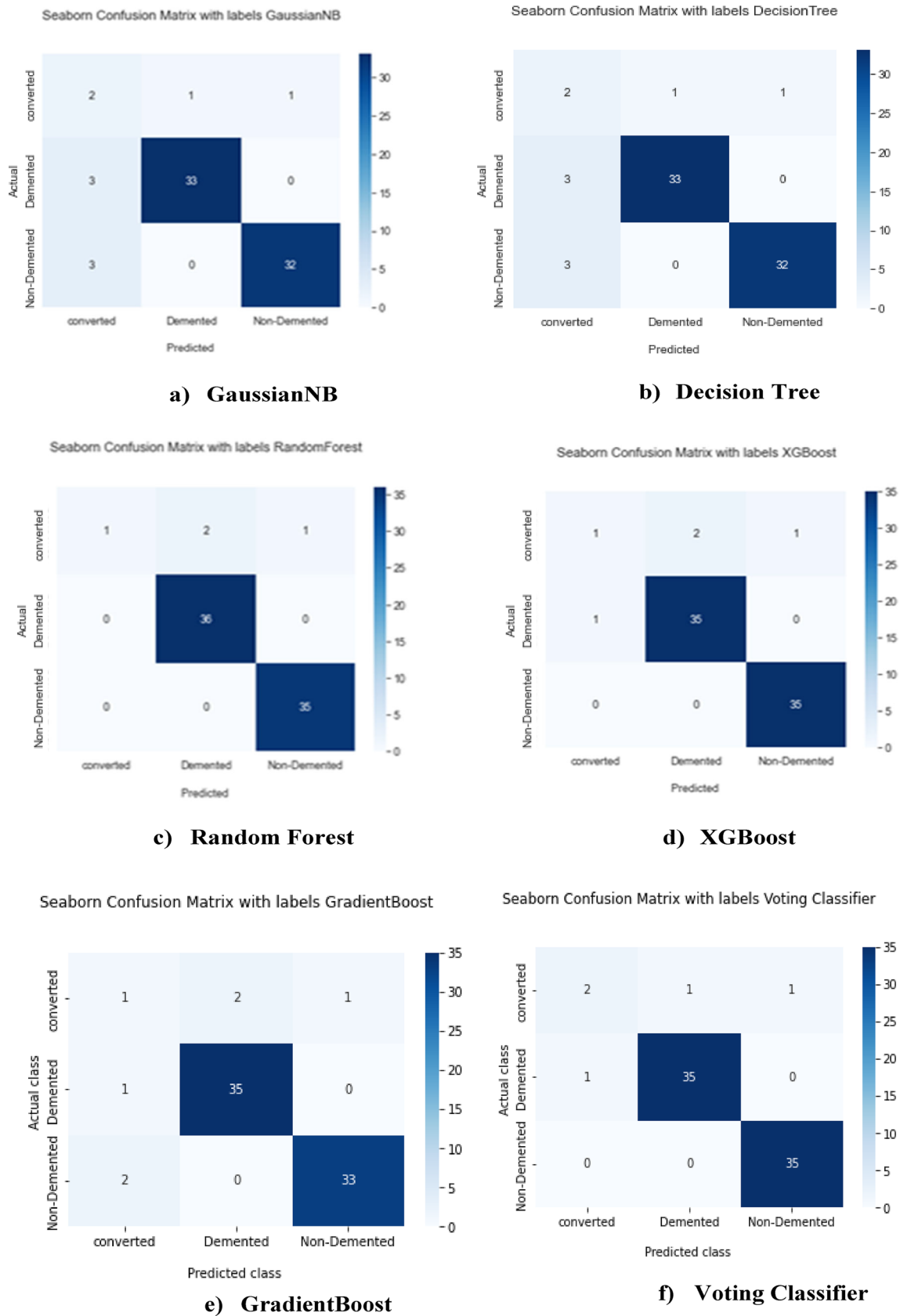


Fig. 10 Confusion matrix for algorithms **a** GaussianNB, **b** decision Tree, **c** random Forest, **d** XGBoost, **e** GradientBoost, **f** voting Classifier

Table 4 Performance evaluation of several ML models

Model	Accuracy (%)	Precision	Recall	F1-score
GaussianNB	96.00	97.00	91.00	94.00
Decision Tree	89.33	97.00	91.00	94.00
Random forest	96.00	97.00	94.00	96.00
Voting classifier	96.00	100.0	43.00	60.00
Xgboost	93.33	97.00	97.00	97.00
GradientBoost	92.00	97.00	94.00	96.00

Data Integrity Issue

Data preprocessing is a crucial stage in machine learning since this helps cleanse, transform, and prepare data for training machine learning models. Nonetheless, preprocessing can possibly compromise the data's integrity, which can have a detrimental impact on the performance of machine learning models for identifying different diseases. However, we have adopted some approaches that can address this issue:

- We apply data augmentation that involves producing new synthetic data from an existing dataset. This can assist in enhancing the diversity and quantity of the dataset, hence, enhancing the robustness of machine learning models.
- We apply cross-validation that is a method for assessing the performance of machine learning models. This includes partitioning the dataset into numerous folds, with each fold serving as a validation set while the remaining folds are used for training. This can assist in identifying problems with the data preprocessing, such as overfitting or underfitting.

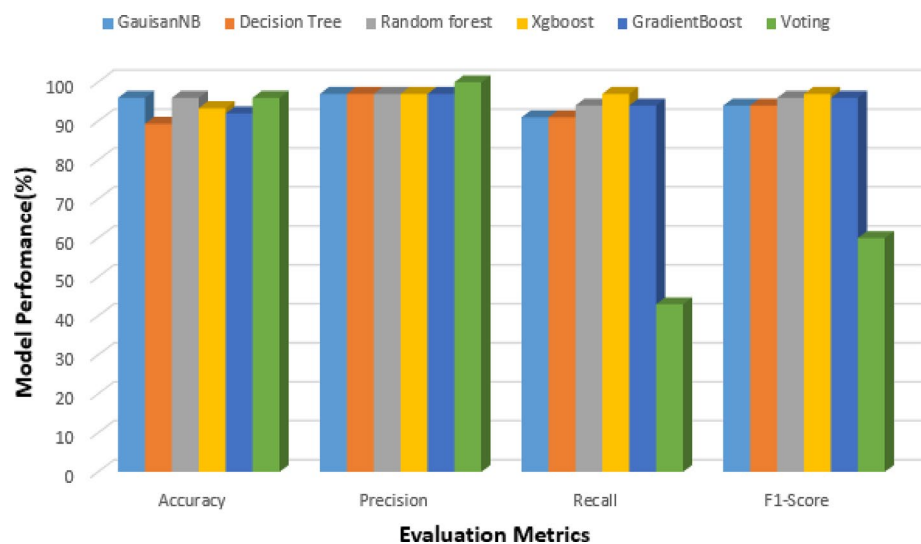
- We have employed feature selection that entails selecting the most pertinent characteristics from a dataset. This can assist lower the dataset's dimensionality and enhance the performance of machine learning models.
- We find out the outliers which are data points that drastically vary from the usual distribution of data.
- Finally, we have developed a robust algorithms that are less susceptible to data preparation problems like missing values, outliers, and skewed distributions.

To overcome the potential integrity issues created by data preprocessing, we apply data augmentation, cross-validation, feature selection, outlier detection, and removal for building a robust AD detection model.

Conclusions

Since presently there is no proven cure for Alzheimer's disease, it is far more important than ever to reduce risk, provide early diagnosis, and thoroughly assess symptoms. The literature research reveals that numerous efforts have been attempted to identify Alzheimer's disease using variety of machine learning algorithms and micro-simulation approaches; nevertheless; it is still daunting to establish relevant traits that might detect Alzheimer's disease at an early stage. In order to identify the most reliable factor for Alzheimer's disease prediction, a number of approaches, including GaussianNB, Decision Tree, Random Forest, XGBoost, Voting Classifier, and GradientBoost, have been used in this study. Feature selection and feature scaling are used in this work to improve the accuracy of the machine learning algorithms. The proposed study yields a more beneficial outcome, with the voting classifier having the greatest validation accuracy of 96% on the test

Fig. 11 Graphical comparison of evaluation metrics attained by the proposed system and existing classifiers



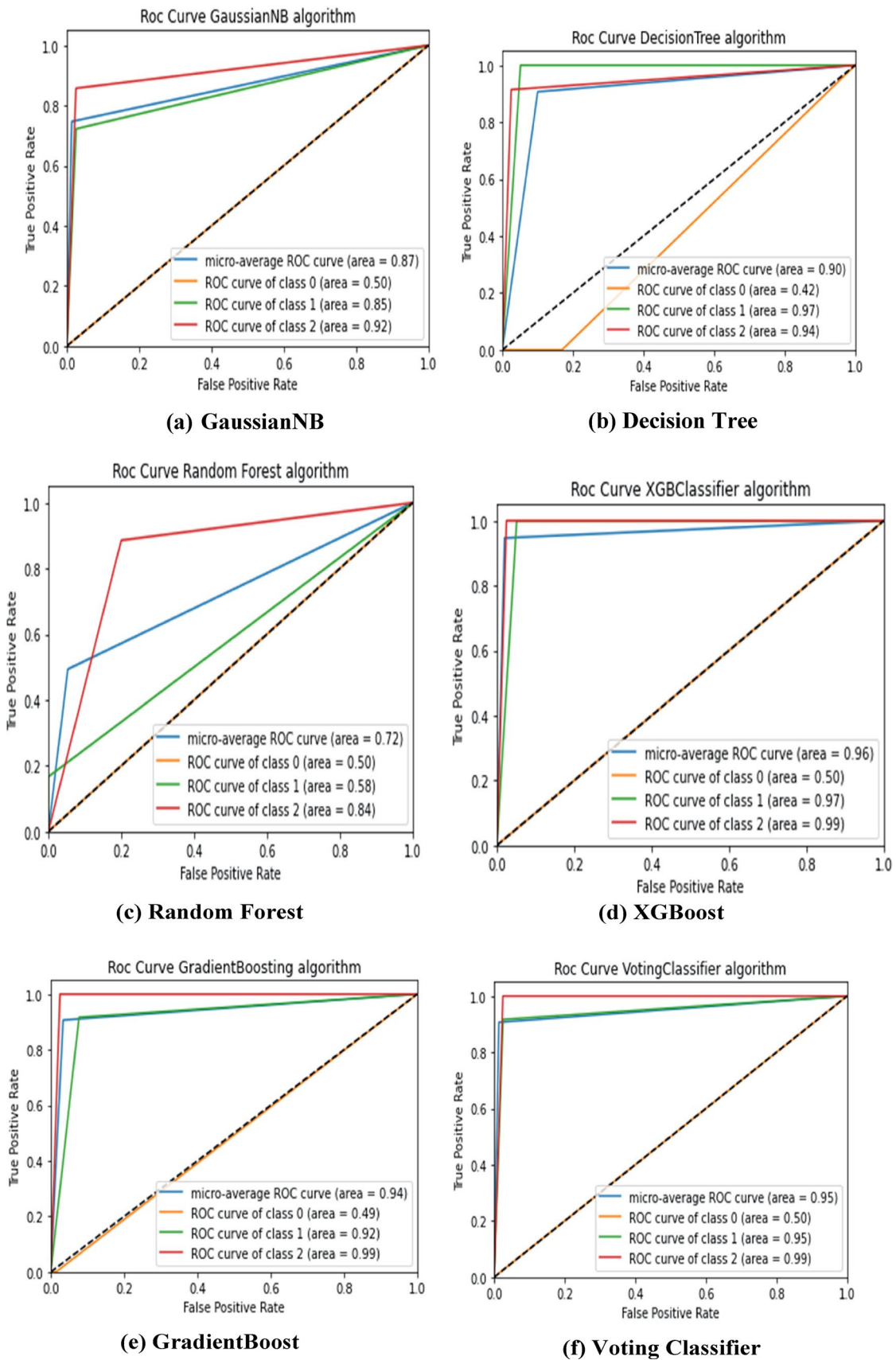


Fig. 12 ROC curve **a** GaussianNB, **b** decision Tree, **c** random Forest, **d** XGBoost, **e** GradientBoost, **f** voting Classifier

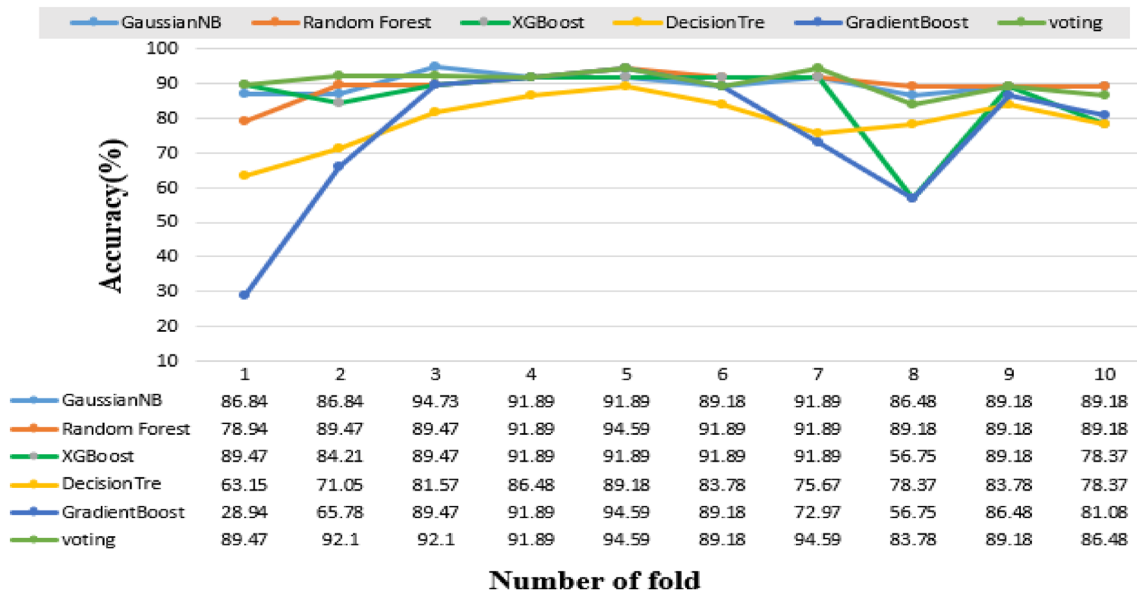


Fig. 13 Comparison of the achieved accuracy of different classifiers

Table 5 Analysis of the proposed model in comparison to the closest-related works

Research study	Dataset	Proposed Method	Classification accuracy (%)
Martinez-Murcia et al. [25]	ADNI dataset	Deep learning (convolutional autoencoders)	80
Helaly et al. [26]	ADNI dataset	Deep learning (VGG19)	97
Kavitha et al. [27]	Imaging Studies (OASIS) data	Machine learning (Majority-based voting)	83
Taher et al. [28]	ADNI dataset	Deep learning(ADDTLA)	91.70
audiuso et al. [29]	Unknown	Laser-Induced Breakdown Spectroscopy (LIBS) and machine learning	80
Basheer et al. [31]	OASIS dataset	Machine learning and deep learning models(M-CapNet)	92.39
Prajapati et al. [32]	ADNI dataset	Deep neural network binary classifier(DNN)	85.19
Lucas et al. [33]	Unknown	Deep feature-based model out performed handcrafted and deep learning(SVM)	79.9
Sudharsan et al. [34]	ADNI dataset	Machine learning technique(SVM) and PCA	78.31
Lin liu et al. [49]	VSBD	Spectrogram feature from audio segmentation (SVC)	83.03
Ortiz et al. [50]	ADNI dataset	Spare features of data (SVC)	92
Our proposed model	OASIS dataset	Machine learning (Voting)	96

data of AD. To improve the detection approaches' accuracy, future research will focus on removing redundant and unneeded characteristics from existing feature sets as well as on extracting and analyzing unique features that are more likely to aid in the detection of Alzheimer's disease.

Author Contributions KMMU: Conceptualization, Methodology, Software, Resources, Writing—Original Draft, Visualization. MJA: Data Curation, Writing—Original Draft. JEAr: Writing—Review & Editing, Visualization. MAU: Writing—Review & Editing. SA: Writing—Review & Editing.

Funding None.

Data Availability The data supporting this study's findings are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that there are no conflicts of interest.

Ethical Approval Not required.

Consent to Participate Not required.

References

1. P. Scheltens, K. Blennow, M.M. Breteler, B. De Strooper, G.B. Frisoni, S. Salloway, W.M. Van der Flier, Alzheimer's disease. *Lancet (Lond, Engl)* **388**(10043), 505–517 (2016)
2. Alzheimer's Disease Facts and Figures. <https://www.alz.org/alzheimers-dementia/facts-figures>. Accessed 10 Aug 2022
3. S. Al-Shoukry, T.H. Rassem, N.M. Makbol, Alzheimer's diseases detection by using deep learning algorithms: a mini-review. *IEEE Access* **8**, 77131–77141 (2020)
4. I.O. Korolev, Alzheimer's disease: a clinical and basic science review. *Med. Student Res. J.* **4**(1), 24–33 (2014)
5. X. Liu, K. Chen, T. Wu, D. Weidman, F. Lure, J. Li, Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease. *Transl. Res.* **194**, 56–67 (2018)
6. D.L. Johnson, R.P. Kesner, Comparison of temporal order memory in early and middle stage Alzheimer's disease. *J. Clin. Exp. Neuropsychol.* **19**(1), 83–100 (1997)
7. J. Taeho, K. Nho, A.J. Saykin, Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. *Front. Aging Neurosci.* **11**, 220 (2019)
8. J. Wen et al., Reproducible evaluation of diffusion MRI features for automatic classification of patients with Alzheimer's disease. *Neuroinformatics* **19**, 57–78 (2021)
9. M. Kivipelto et al., World-Wide FINGERS Network: a global approach to risk reduction and prevention of dementia. *Alzheimers Dement.* **16**(7), 1078–1094 (2020)
10. J. Bona et al., Semantic integration of multi-modal data and derived neuroimaging results using the platform for imaging in precision medicine (PRISM) in the Arkansas imaging enterprise system (ARIES). *Front. Artif. Intell.* **4**, 649970 (2022)
11. M. Van Zeller, D. Dias, A.M. Sebastião, C.A. Valente, NLRP3 inflammasome: a starring role in amyloid- β -and tau-driven pathological events in Alzheimer's disease. *J. Alzheimers Dis.* **83**(3), 939–961 (2021)
12. R. Mostafiz, M.S. Uddin, K.M.M. Uddin, M.M. Rahman, COVID-19 along with other chest infections diagnosis using faster R-CNN and generative adversarial network. *ACM Trans. Spatial Algorithms Syst.* (2022). <https://doi.org/10.1145/3520125>
13. R. Hertel, R. Benlamri, A deep learning segmentation-classification pipeline for X-ray-based covid-19 diagnosis. *Biomed. Eng. Adv.* **3**, 100041 (2022)
14. A. Chattopadhyay, M. Maitra, MRI-based brain tumor image detection using CNN based deep learning method. *Neurosci. Inform.* **2**, 100060 (2022)
15. S.K. Mamatha, H.K. Krishnappa, N. Shalini, Graph theory based segmentation of magnetic resonance images for brain tumor detection. *Pattern Recogn. Image Anal.* **32**(1), 153–161 (2022)
16. G.N. Ahmad, H. Fatima, A.S. Saidi, Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV. *IEEE Access* **10**, 80151 (2022)
17. M.M. Rahman, M.R. Rana, Nur-A-Alam, M.S.I. Khan, K.M.M. Uddin, A web-based heart disease prediction system using machine learning algorithms. *Netw. Biol.* **12**(2): 64–81 (2022)
18. S.K. Dey, M.M. Rahman, A. Howlader, U.R. Siddiqi, K.M.M. Uddin, R. Borhan, E.U. Rahman, Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data in Bangladesh: a machine learning approach. *PLoS ONE* **17**(7), e0270933 (2022)
19. D.V.B. Oliveira, J.F. da Silva, T.A. de Sousa-Araújo, U.P. Albuquerque, Influence of religiosity and spirituality on the adoption of behaviors of epidemiological relevance in emerging and re-emerging diseases: the case of dengue fever. *J. Relig. Health* **61**(1), 564–585 (2022)
20. N. Biswas, K.M.M. Uddin, S.T. Rikta, S.K. Dey, A comparative analysis of machine learning classifiers for stroke prediction: a predictive analytics approach. *Healthcare Anal.* **2**, 100116 (2022)
21. H. Liao, R. Fang, J.B. Yang, D.L. Xu, A linguistic belief-based evidential reasoning approach and its application in aiding lung cancer diagnosis. *Knowl.-Based Syst.* **253**, 109559 (2022)
22. J. Akhtar, Non-Small Cell Lung Cancer Classification from Histopathological Images using Feature Fusion and Deep CNN. *Int. J. Eng. Adv. Technol.* (2020). <https://doi.org/10.35940/ijeat.E9266.069520>
23. P. Khan, M.F. Kader, S.R. Islam, A.B. Rahman, M.S. Kamal, M.U. Toha, K.S. Kwak, Machine learning and deep learning approaches for brain disease diagnosis: principles and recent advances. *IEEE Access* **9**, 37622–37655 (2021)
24. A. Mehmood, M. Maqsood, M. Bashir, Y. Shuyuan, A deep Siamese convolution neural network for multi-class classification of Alzheimer disease. *Brain Sci.* **10**(2), 84 (2020)
25. F.J. Martinez-Murcia, A. Ortiz, J.M. Gorriz, J. Ramirez, D. Castillo-Barnes, Studying the manifold structure of Alzheimer's disease: a deep learning approach using convolutional autoencoders. *IEEE J. Biomed. Health Inform.* **24**(1), 17–26 (2019)
26. H.A. Helaly, M. Badawy, A.Y. Haikal, Deep learning approach for early detection of Alzheimer's disease. *Cogn. Comput.* **14**, 1711 (2021)
27. C. Kavitha, V. Mani, S.R. Srividhya, O.I. Khalaf, C.A.T. Romero, Early-stage Alzheimer's disease prediction using machine learning models. *Front. Public Health* (2022). <https://doi.org/10.3389/fpubh.2022.853294>
28. T.M. Ghazal, G. Issa, Alzheimer disease detection empowered with transfer learning. *Comput. Mater. Continua* **70**(3), 5005–5019 (2022)
29. R. Gaudiuso, E. Ewusi-Annan, W. Xia, N. Melikechi, Diagnosis of Alzheimer's disease using laser-induced breakdown spectroscopy and machine learning. *Spectrochim. Acta Part B* **171**, 105931 (2020)
30. H. Nawaz, M. Maqsood, S. Afzal, F. Aadil, I. Mehmood, S. Rho, A deep feature-based real-time system for Alzheimer disease stage detection. *Multimedia Tools Appl.* **80**(28), 35789–35807 (2021)
31. S. Basheer, S. Bhatia, S.B. Sakri, Computational modeling of dementia prediction using deep neural network: analysis on OASIS dataset. *IEEE Access* **9**, 42449–42462 (2021)
32. R. Prajapati, U. Khatri, and G.R. Kwon, An efficient deep neural network binary classifier for Alzheimer's disease classification, in 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC) (pp. 231–234) (IEEE, 2021)
33. L.R. Trambaiolli, A.C. Lorena, F.J. Fraga, P.A. Kanda, R. Anghinah, R. Nitrini, Improving Alzheimer's disease diagnosis with machine learning techniques. *Clin. EEG Neurosci.* **42**(3), 160–165 (2011)
34. M. Sudharsan, G. Thailambal, Alzheimer's disease prediction using machine learning techniques and principal component analysis (PCA). *Mater. Today: Proc.* (2021). <https://doi.org/10.1016/j.matpr.2021.03.061>
35. O.I. Khalaf, G.M. Abdulsahib, Frequency estimation by the method of minimum mean squared error and P-value distributed in the wireless sensor network. *J. Inf. Sci. Eng.* **35**(5), 1099–1112 (2019)
36. AD Dataset, <https://www.kaggle.com/datasets/shashwatwork/dementia-prediction-dataset>. Accessed 10 Aug 2022
37. A.R. Javed, L.G. Fahad, A.A. Farhan, S. Abbas, G. Srivastava, R.M. Parizi, M.S. Khan, Automated cognitive health assessment in smart homes using machine learning. *Sustain. Cities Soc.* **65**, 102572 (2021)
38. C.L. Saratxaga, I. Moya, A. Picón, M. Acosta, A. Moreno-Fernandez-de-Leceta, E. Garrote, A. Bereciartua-Perez, MRI deep

- learning-based solution for Alzheimer's disease prediction. *J. Personaliz. Med.* **11**(9), 902 (2021)
39. T.R. Gadekallu, C. Iwendi, C. Wei, Q. Xin, Identification of malnutrition and prediction of BMI from facial images using real-time image processing and machine learning. *IET Image Process* **16**, 647–658 (2021)
 40. K. Ota, N. Oishi, K. Ito, H. Fukuyama, Sead-J Study Group and Alzheimer's Disease Neuroimaging Initiative, Effects of imaging modalities, brain atlases and feature selection on prediction of Alzheimer's disease. *J. Neurosci. Methods* **256**, 168–183 (2015)
 41. J. Wan, Z. Zhang, B.D. Rao, S. Fang, J. Yan, A.J. Saykin, L. Shen, Identifying the neuroanatomical basis of cognitive impairment in Alzheimer's disease by correlation-and nonlinearity-aware sparse Bayesian learning. *IEEE Trans. Med. Imaging* **33**(7), 1475–1487 (2014)
 42. J.M. Rasmussen, A. Lakatos, T.G. van Erp, F. Kruggel, D.B. Keator, J.T. Fallon, F. Maciardi, S.G. Potkin, Empirical derivation of the reference region for computing diagnostic sensitive 18fluoro-deoxyglucose ratios in Alzheimer's disease based on the ADNI sample. *Biochim. Biophys. Acta (BBA) Mol. Basis Dis.* **1822**(3), 457–466 (2012)
 43. H. Kamel, D. Abdulah, J.M. Al-Tuwaijari, Cancer classification using gaussian naive bayes algorithm, in 2019 International Engineering Conference (IEC) (pp. 165–170) (IEEE, 2019)
 44. T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in Proceedings of the 22nd acmsigkdd International Conference on Knowledge Discovery and DATA mining (pp. 785–794) (2016)
 45. A.D. Dana, A. Alashqur, Using decision tree classification to assist in the prediction of Alzheimer's disease, in 2014 6th International Conference on Computer Science and Information Technology (CSIT) (pp. 122–126) (IEEE, 2014)
 46. N.K Dewi, U.D Syafitri, S.Y. Mulyadi, Penerapan Metode Random Forest dalam Driver Analysis, in Forum Statistika dan Komputasi (Vol. 16, No. 1) (2011)
 47. L.V. Fulton, D. Dolezel, J. Harrop, Y. Yan, C.P. Fulton, Classification of Alzheimer's disease with and without imagery using gradient boosted machines and ResNet-50. *Brain Sci.* **9**(9), 212 (2019)
 48. A.S. Assiri, S. Nazir, S.A. Velastin, Breast tumor classification using an ensemble machine learning method. *J. Imaging* **6**(6), 39 (2020)
 49. L. Liu, S. Zhao, H. Chen, A. Wang, A new machine learning method for identifying Alzheimer's disease. *Simul. Model. Pract. Theory* **99**, 102023 (2020)
 50. A. Ortiz, F. Lozano, J.M. Gorriz, J. Ramirez, F.J. Martinez Murcia, Alzheimer's Disease Neuroimaging Initiative, Discriminative sparse features for Alzheimer's disease diagnosis using multimodal image data. *Curr. Alzheimer Res.* **15**(1), 67–79 (2018)

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.