Annals of Data Science
https://doi.org/10.1007/s40745-022-00452-2

Performances of Machine Learning Models for Diagnosis of Alzheimer's Disease

Siddhartha Kumar Arjaria[1] · Abhishek Singh Rathore[2] · Dhananjay Bisen[3] · Sanjib Bhattacharyya[4]

Received: 29 November 2021 / Revised: 18 April 2022 / Accepted: 7 May 2022

Abstract
In recent times, various machine learning approaches have been widely employed for effective diagnosis and prediction of diseases like cancer, thyroid, Covid-19, etc. Likewise, Alzheimer's (AD) is also one progressive malady that destroys memory and cognitive function over time. Unfortunately, there are no dedicated AI-based solutions for diagnoses of AD to go hand in hand with medical diagnosis, even though multiple factors contribute to the diagnosis, making AI a very viable supplementary diagnostic solution. This paper reports an endeavor to apply various machine learning algorithms like SGD, k-Nearest Neighbors, Logistic Regression, Decision tree, Random Forest, AdaBoost, Neural Network, SVM, and Naïve Bayes on the dataset of affected victims to diagnose Alzheimer's disease. Longitudinal collections of subjects from OASIS dataset have been used for prediction. Moreover, some feature selection and dimension reduction methods like Information Gain, Information Gain Ratio, Gini index, Chi-Squared, and PCA are applied to rank different factors and identify the optimum number of factors from the dataset for disease diagnosis. Furthermore, performance

✉ Siddhartha Kumar Arjaria
arjarias@gmail.com

Abhishek Singh Rathore
abhishekatujjain@gmail.com

Dhananjay Bisen
bisen.it2007@gmail.com

Sanjib Bhattacharyya
sanjiv2k1@yahoo.co.in

1 Rajkiya Engineering College, Banda, India

2 Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India

3 Department of Information Technology, Madhav Institute of Technology and Science, Gwalior, India

4 Department of Pharmaceutical Science and Chinese Traditional Medicine, Southwest University, Chongqing, China

Published online: 17 October 2022

is evaluated of each classifier in terms of ROC-AUC, accuracy, F1 score, recall, and precision as well as included comparative analysis between algorithms. Our study suggests that approximately 90% classification accuracy is observed under top-rated four features CDR, SES, nWBV, and EDUC.

Keywords Alzheimer's disease · Data science · Machine learning · Feature selection · Classification algorithms

1 Introduction

In the modern competitive world, the lifestyle of the person is hustled, irregular, and stretched towards multitasking, resulting in many health-related issues leading to long-term chronic diseases. Alzheimer's disease (AD) belongs to the category of diseases associated with progressive dementia that causes impairment of memory, thought process, and conduct [1]. Symptoms typically grow gradually and deteriorate over the time, eventually making it impossible for the patient to perform even basic everyday tasks, as the patient gradually loses ability to control their body, and in cases, can lead to death [2]. Alzheimer's disease (AD) is incurable; hence, early detection and diagnosis are very important to achieve better disease management. Conventional indicative and demonstrative techniques for AD detection have a few restrictions, and they depend essentially on certain scales, clinical presentation or CT, MRI, and so forth. For instance, numerous components can meddle with the Mini-mental State Examination. But these measures are tedious and hard to apply to patients in serious infectious stages of AD. CT-Scan and MRI on the other hand are costly and hence not affordable to screen sicknesses in a huge populace. Even though, several computer-aided initiatives have been developed in recent years to help in the early detection of diseases; their cost makes their widespread application very difficult, especially in poor economies [3].

Data science based approaches are intuitive and can learn from experience, and hence can be complementary to medical diagnosis and thus are used extensively in healthcare sector [4] which is one of the key indicator of a countries human development index [5]. With high availability of digital pathology dataset, and recent advancement in predictive data analytical methods, various data science approaches have been widely used to diagnose diseases with complicated etiologies. Medical resonance imaging (MRI) is a non-invasive technique, which is comparatively cheaper than invasive methods, and is easier to understand from an analytical point of view as it captures brain microstructure with a high spatial resolution. Hence, MRI has been used to generate data regarding informative biomarkers for mapping Alzheimer's disease [1, 2]. Recent advancements in data acquisition and analysis techniques have made it possible to analyze the volumes of medical data. Further, advancements and availability of these tools has made it possible for interdisciplinary cooperation between domain experts, ultimately leading to improvements in medical analysis and patient care. This paper provides a detailed study of predictive data science algorithms that contribute to the field of early and efficient diagnosis of Alzheimer's disease. The goal of early AD prediction is to perform predictive analytics and evaluate the effectiveness

of standard machine learning (ML) algorithms that provides insights to researchers to infer patients' status and will be further used in prescriptive analytics for decision making. The feature selection techniques applied in this work, is used provide more valued information for AD diagnosis. Zhang et al. [6] focused on more informative features and thus applied the feature selection method. P-order L_2 norm regularization was applied for feature selection and thus having the advantage of robust performance over L_2 norms. Although the method can be used for non-smooth optimization problems it was unable to find correlations for multi-views. Likewise, Zhang et al. [7] applied multi-view classification on the MRI dataset by considering intra-structure and the inter-structure relationship between features with $L_{2,p}$ norms. Luo et al. [8] created a feature subset using SVM-RFE, mRMR [9], and RF feature selection and applied SVM to handle the small size of datasets with high dimensionality.

Consequently, this paper includes various classifiers to diagnose Alzheimer's disease that is based on machine learning algorithms like SGD, k-Nearest Neighbors, Logistic Regression, Decision tree, Random Forest, AdaBoost, Neural Network, SVM, and Naïve Bayes. Historically and geographically, AD has been known to affect people who are above 65 years of age. Therefore, the dataset used in this study covers subjects of age 60 to 90 years old. Afterward, feature selection and dimension reduction methods like Gini index, Information Gain, Information Gain Ratio, Chi-Squared, and PCA are applied to identify the optimum number of features. Finally, performance of each classifier is evaluated in terms of ROC-AUC curve, accuracy, F1 Score, recall, and precision.

The rest of the paper is organized as follows. Section 2 covers the related work that shows researchers' contribution in the field of early and efficient Alzheimer's disease diagnosis using Machine Learning algorithms. Section 3 examined different classification models and datasets in detail. Section 4 presents the experimental result and analysis with a comparison from the current state of the art, while Sect. 5 concludes this paper.

2 Literature Review

Alzheimer's is characterized by sudden and rapid deterioration in the mental faculty of people suffering from it hence, there is an urgent need for early diagnosis of AD to provide cost-effective solutions for disease management in the form of medication to patients in early stages of AD. This section includes the existing work done in the field of early predictions of AD. Mainly, the research has been focused in two different areas, feature selection, and classification tasks. Some authors used neuroimaging datasets while others used the EEG dataset. Almost in all cases, researchers focused on a binary classification task. In machine learning-based algorithms, data works as nutrients, and results may vary with the size of the datasets available. Because of the small sample size and the high dimension of small datasets, Zheng et al. [10] used MRBM as the LUPI algorithm for feature learning from MRI images and considered SVM and SVM+ as an ensemble for early diagnosis of AD. The integrated RBM and SVM+ models worked considerably well on small datasets. Afterward, different CNN architectures were designed by Ji et al. [11], while Valliani et al. [12] used deep residual CNN,

and Cherdal and Mouline [13] created Petri Nets for modeling the complex structure of the brain. McCrackin [14] applied deep CNN with data augmentation based on extrapolation and interpolation to achieve better accuracy. Furthermore, Liu et al. [15] applied multi-class classification on MRI using deep CNN. Additionally, the authors have used regression analysis with a deep neural network to find the decaying rate of the patient's brain. Even though the model performs well, it lacks in robustness with observed variations in the distribution of data.

With a small dataset of 28 people, Zhao and He [16] combined a deep learning model with SVM using incremental learning for early AD prediction. A voxel-based morphometry (VBM) [1] approach to ensemble PCA-based bagging and boosting is applied on the MRI dataset. No prior information is required for using VBM even for a small dataset, but the complexity of classification is higher in ensemble classifications. In the voxel-based analysis, often generalized linear model/logistic regression is applied to predict disease and the desired estimator along with graph difference operator is sparse. But there exists a procedural bias that violates the property of prior sparsity. Consequently, Sun et al. [17] applied variable splitting with generalized linear models to gain better predictions, but with unimodal data. To capture more complex relationships Cao et al. [18] capture non-linear relationships between features and response variables with preserving sparsity. Liu et al. [19] learned cognitive features using the Laplacian sparse group lasso model while maintaining sparsity. With this multitask learning model, cognitive measure over time is still an open issue that can improve the prediction power of machine learning models. To maintain sparsity, Xu et al. [20] applied multi-task learning with dual margin loss function of behavioral and background data of patients.

To diagnose AD early, mild cognitive impairment (MCI) can be considered as early-stage where patients over time remain stable or change to AD. Zhang et al. [21] predict the conversion of MCI to AD using semi-supervised learning with Laplacian SVM. Accordingly, Zhu et al. [22] focused on incomplete multimodal data to identify the label-data relationship for classification tasks. The multimodal data is assumed to be a mixture of multiple distributions and Maximum Mean Discrepancy based distance is used to find the difference between the distributions. SVM is applied to the estimated distance to classify the MRI images. The generalized classifier works well on homogenous data, but clinical data shows heterogeneous characteristics due to various complex distributions. Zhu et al. [23] capture distribution divergence in the high dimension space with biconvex optimization for designing a personalized classifier. While Gamberger [24] applied gender-based clustering, an unsupervised ML approach to a group of patients, thus making homogenous groups of patients for further application of ML algorithms. Li [25] focused on the open challenges of i.e. AD staging and avoiding retraining of classifiers. AD staging is optimized by α expansion while the bottleneck of retraining is solved by 3 steps Multifold Bayesian Kernalizaiton. In addition to MRI and CT, researchers also worked on EEG and MEG for the early diagnosis of AD. Abasolo et al. [26] applied sample entropy to extract relevant features from EEG data. Although the size of sample entropy does not depend on the size of the data set, (samples are assumed to be independent), parameter setting is crucial and results vary with change in parameter setting.

From the various studies [27] it has been found that AD may affect brain topological structure, hence, BrainNetCNN [28] is been proposed to learn the topology of the brain. In addition to that Spectral CNN is developed to learn the remaining geometric feature in Euclidean Space, giving better feature selection than BrainNetCNN. On the other hand, Palafox [29] applied a mean shift algorithm for the segmentation of the hippocampus of MRI, using SVM and kernel functions to map changes that came out in the brain, starting with AD. Chen et al. [30] applied Gaussian Probability-based segmentation for identifying the decaying structure of the brain. The two open challenges with almost all models are, first the small size of the data set with high dimensionality and second is the training of classifier with new datasets. Consequently, this paper considers all issues and introduces various classifiers to diagnose Alzheimer's disease that is based on machine learning algorithms. Feature selection and dimension reduction methods are also applied to give rank to features and identify the optimum number of features, and each classifier is evaluated in terms of ROC-AUC curve, accuracy, F1 Score, recall, and precision.

3 Dataset and Methods

This article includes different methods and datasets used for the performance evaluation of different ML algorithms. The work is broadly echeloned into two parts. First is the feature selection method, where the focus is to get more relevant features and discard the extraneous for faster and more efficient classification of data. In the second part, the classification algorithms are applied to obtained features to make predictions.

3.1 Data Set Description

The main problem with existing work is the small size of datasets available, $D \in \mathbb{R}^{n \times m}$ where, $n \approx m$, i.e. small sample size with a large number of features, and therefore approximating the parameters are difficult. The Oasis-2 Brain Data Set [31] is used for the early diagnosis of AD. It contains MRI scans of 150 subjects with 373 sessions. Each subject's age ranges from 60 to 96 is shown in Fig. 1 in the form of a histogram in which red and blue stripe represents male and female count (frequency) respectively of a particular age. The data is classified into 3 classes, demented, non-demented, and converted. The distribution of data is depicted in Fig. 2 and Table 1. Here, the x-axis and y-axis show the group and frequency respectively, it denotes male and female count under each group in the dataset. In this manner, the dataset is suitable for molding the organization loads for accomplishing the most ideal arrangement of results.

3.2 Feature Selection Techniques

The dataset contains 13 attributes and 3 classes. The distribution of all three classes is depicted in Fig. 3. In the decision-making process, including all features makes computational costs higher. Therefore, feature selection methods are used for selecting

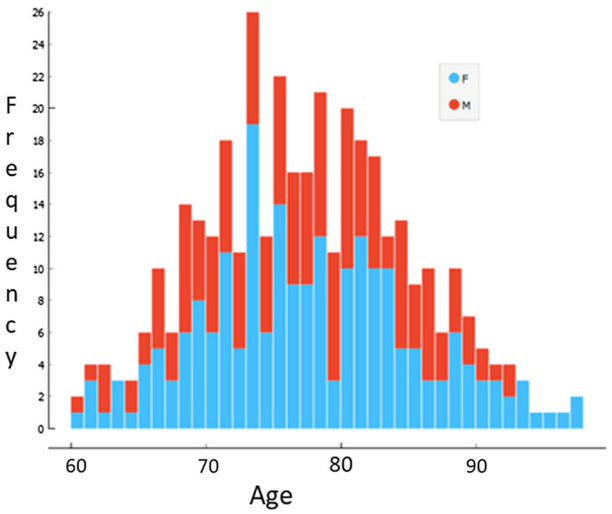

Fig. 1 The age-wise distribution of data

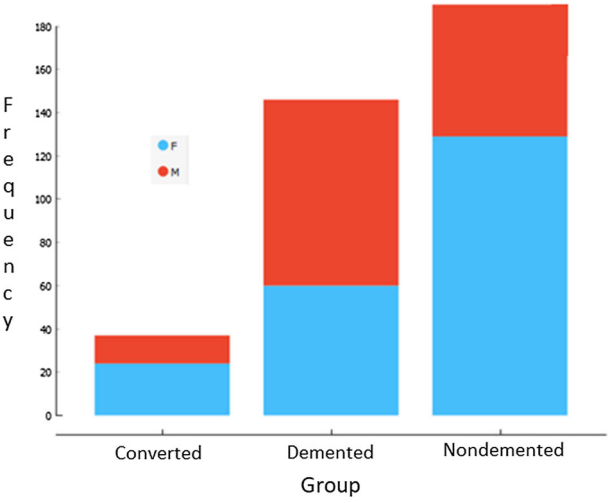

Fig. 2 The distribution of data with 3 different classes

Table 1 Distribution of data

Class/Count	F	M	Total
Converted	24	13	37
Demented	60	86	146
Nondemented	129	61	190
Total	213	160	373

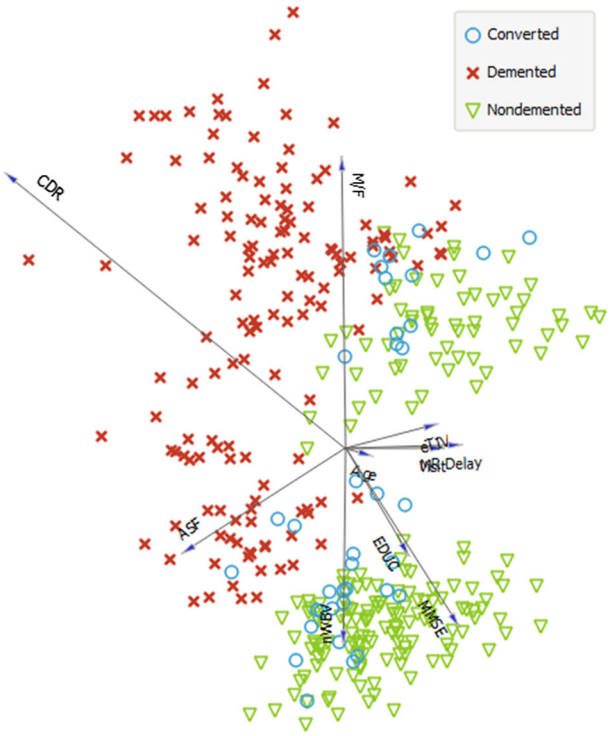

Fig. 3 The feature space to the data

relevant features with some loss of information. To identify the optimum number of features following techniques namely Information-Gain, Information-Gain Ratio, Gini-Index, Chi-squared, and a dimension reduction technique (Principal Component Analysis) is applied to select relevant features from the data.

The reduction in the entropy is measured as Information gain [32]. It takes the product of probabilities of class with a log having base 2 of that class probability. Entropy measures the minimum amount of information needed to classify the data. Thus, information gain gives the impact of selecting some of the features from all available features. Higher the information gain, the better the purity gained on splitting data based on a few features. Initially calculate Entropy to measure the purity of split and then information gain (IG) to determine which feature gives us the maximum information about a class, shown in Eqs. (1) and (2):

$$Entropy\ H(X) = -\sum_{i=1}^{N} P_i * \log_2 P_i \tag{1}$$

where X, N and P_i represent the set of all instances in the dataset, number of distinct class values and even probability respectively.

$$IG(A, X) = H(X) - H(A, X) = \sum_{y=1}^{v} \left| \frac{X_j}{X} \right| * H(X_j) \tag{2}$$

where H(X) Entropy of dataset X, $|X_j|$ number of instances with y value of an attribute A, |X| total number of instances in dataset X, v set of distinct value of an attribute A, $H(X_j)$ entropy of subset of instance for attribute A, H(A,X) entropy of an attribute A. The information gain is biased towards those attributes which have where high variability in data values. To remove biased for such cases, Information Gain Ratio is calculated that is divided by its intrinsic value i.e. the conditional entropy.

Information gain covers lesser distribution while calculating which feature is prominent in decision making. To incorporate higher value of distributions Gini index is used, which is calculated by Eq. (3). It captures the variance of distributions that are associated with features. Its value lies between 0 to 1, where 0 represents equality and 1 represents inequality [33]. A Gini index of 0.5 denotes equally distributed elements into some classes. For feature selection, a higher value of Gini-index shows that it has more independent features in the dataset.

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2 \tag{3}$$

Other methods like chi-squared (X^2) [8, 9] aim at selecting optimal features or finding relation between features within the dataset, and are used to measure the association between features with the n − 1 degree for freedom. More independent features will contribute more to decision making and hence those features are more relevant. In this method, firstly null and alternate hypothesis defines to check the correlation between features, whether there is a significant relation between features or not. Here, the significance level is 0.05 and its chi-square tabular value with a degree of freedom (shown in Eq. 4) is 21.03, meanwhile, Chi-square is calculated by Eq. (5):

$$Degree\ of\ \ freedom = (columns - 1)(rows - 1) \tag{4}$$

$$X^2_{Calculated} = \sum \frac{(Observed\ value - Expected\ value)^2}{Expected\ value} \tag{5}$$

Another technique is Principal Component Analysis (PCA) [1], which is a dimension reduction technique. It transforms the data into new smaller feature space and still contains most of the information of original data. It captures the total variability covered by identified Principal components as depicted in Fig. 4. It can be seen from the figure that the newly transformed feature space with 6 principal components covers 75% variability of overall data.

Finally, Table 2 depicts the comparative score features by applying the above four mentioned feature selection techniques. All the techniques used gave almost the same results with CDR having the highest contribution while Hand (Right-Handed or Left) is the irrelevant feature. These results are comparable with the visualization of feature

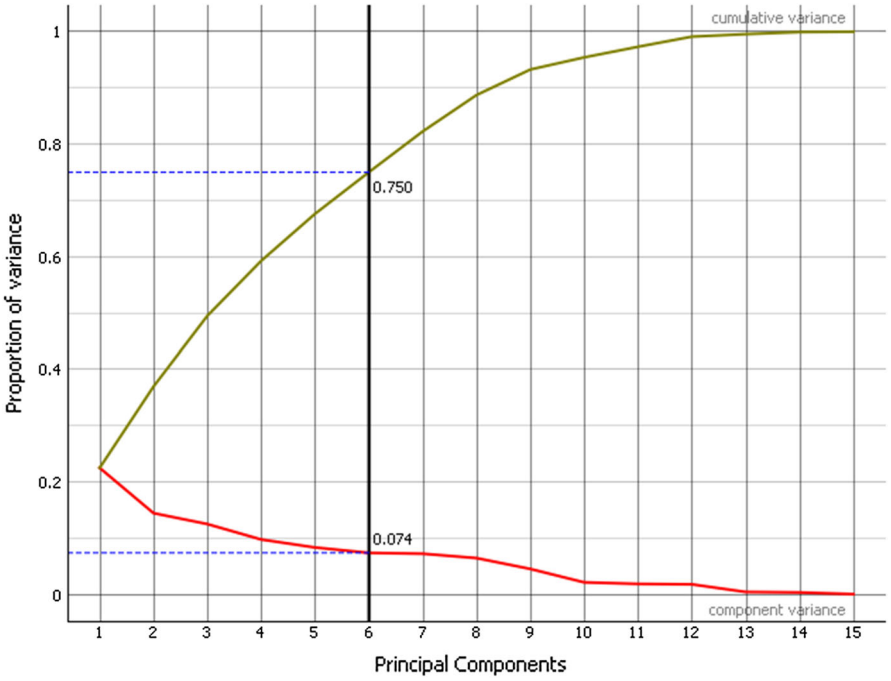

Fig. 4 Variability covered by the number of principal components

space obtained in Fig. 3. Therefore, it can be stated that all the methods are unbiased on the current data set and gives the same ordering of the features except the Gini Index, since Gini Index captures distribution variance, but results are near about the same. CDR, MMSE, SES, nWBV, and EDUC have higher scores and contribute more than other features in decision making.

3.3 Classification Techniques

In classification technique, the dataset is usually divided into training and testing set in which classifier learns from the training set (features of data) at the time of training and then evaluate itself using the unseen test set. Finally, it predicts the precise class label for a given input. In this work, Stochastic Gradient Descent, k-Nearest Neighbors, Logistic Regression, Decision tree, Random Forest, AdaBoost, SVM, Neural Network have been used for classification.

Stochastic Gradient Descent [34] is an optimization method that defines the optimal cost/loss function for a problem. It learns by considering a single training data at a time and tries to find the coefficient of function under the condition that minimizes the loss margin (J(Q)). The algorithm repeats over the training samples, that updates the model parameters (weights $Q(\beta_j)$ $(\beta_0, \beta_1, \ldots)$) corresponding to each training example, the

Table 2 Scores obtained by applying feature selection techniques

Features	Info. gain	Gain ratio	Gini	χ^2
CDR	0.877	0.623	0.394	247.732
MMSE	0.413	0.208	0.212	134.756
SES	0.080	0.038	0.030	21.415
nWBV	0.073	0.036	0.042	29.116
EDUC	0.067	0.034	0.036	28.009
MR delay	0.058	0.030	0.026	14.759
M/F	0.049	0.050	0.029	14.399
Age	0.023	0.012	0.007	5.996
Visit	0.018	0.011	0.009	5.523
eTIV	0.012	0.006	0.003	0.602
ASF	0.011	0.006	0.003	0.752
Hand	0.000	0.000	0.000	Nan

update rule is given by Eqs. (5) and (6) and try to minimize the sum of squared errors.

$$J(Q) = \frac{1}{2}\sum_{i=1}^{m}(h(X^i) - Y^i)^2$$

$$h(X^i) = \sum_{i=0}^{m}\beta_i x_i \quad \forall\, training\ examples\ (m)$$

The weight (Q) is updated continuously to make J(Q) smaller, until it converges to global minima.For a single training sample, the update delta rule is:

$$Q_j = Q_j - \propto \frac{\partial}{\partial Q_j} J(Q)$$

$$\beta_j = \beta_j + \propto (Y^i - h(X^i))X^i_j \tag{5}$$

From Eq. (5), it is observed that if Y and h(X) both are the same then β_j do not change and if Y is greater than h(X), the value of β_j must be increased therefore h(X) comes closer to Y, likewise β_j will be updated for each training example and repeated until convergence.

Repeat {
  for i = 1 to m do
  $Q_j = Q_j + \propto (Y^i - h(X^i))X^i_j$ (for every j) (6)
  end for
} until convergence

Table 3 SGD parameters

S. no	Parameters	Value
1	Classification loss function	Hinge
2	Regression loss function	Squared loss
3	Regularization	Ridge (L2)
4	Regularization strength (α)	1e−05
5	Learning rate	Constant
6	Initial learning rate (η0)	0.01
7	Shuffle data after each iteration	Yes

where Y^i is the actual value for a particular sample (x^i, y^i), $h(X^i)$ is the estimated value and $\propto$ is the learning rate. In this analysis, parameters have considered under SGD shown in Table 3.

Another method is K-Nearest Neighbor [35], which is a straightforward, easy-to-implement supervised machine learning algorithm that is used to solve classification and regression problems. KNN assumes that similar things exist close and captures the concept of similarity (sometimes referred to as distance, proximity, or closeness), using some arithmetic operation and shows the gap between points on a graph using Euclidean distance formula, illustrated in Eq. (7). To pick out the K that's right for the information, algorithm is evaluated multiple times with random K values and optimized to select the K that reduces the number of errors. KNN's main disadvantage is that it becomes considerably slower as the volume of information increases thus, making it an impractical choice in environments wherever predictions ought to be made rapidly. KNN classifier consider uniform weight, Euclidean metric, and several neighbors 5 to detect Alzheimer's disease.

$$d(i,j) = \sqrt{|X_{i_1} - X_{j_1}|^2 + |X_{i_2} - X_{j_2}|^2 + |X_{i_3} - X_{j_3}|^2 + \cdots + |X_{i_n} - X_{j_n}|^2} \tag{7}$$

$$d(i,j) = \sqrt{\sum_{k=1}^{n} (|X_{i_k} - X_{j_k}|)^2}$$

Furthermore, logistic regression [15] is applied to Alzheimer's disease dataset with ridge regularization (L2), it is appropriate regression or predictive analysis and employed to clarify the relation between a dependent binary variable and one or more ordinal features. Logistic Regression is employed once the dependent variable i.e., the target variable is categorical. Mathematically, cost function estimation is shown in Eqs. (8) and (9) under logistic regression is defined as:

$$cost\ function = \max \sum_{i=1}^{n} y_i \times w^t x_i \tag{8}$$

Here w^t updates every time till find out the maximum value of the cost function to get a best-fit line. Here classification is dependent on some conditions such as:

Case 1: if $y_i = +1$ and $w^t x_i > 0$ then $y_i * w^t x_i > 0$, that means the data point is correctly classified.
Case 2: if $y_i = -1$ and $w^t x_i < 0$ then $y_i * w^t x_i > 0$, that means the data point is correctly classified.
Case 3: if $y_i = -1$ and $w^t x_i > 0$ then $y_i * w^t x_i < 0$, that means the data point is not correctly classified.
Case 4: if $y_i = +1$ and $w^t x_i < 0$ then $y_i * w^t x_i < 0$, that means the data point is not correctly classified.

Now correctly classify the outliers under this dataset, the sigmoid function is used, shown in Eq. (10) therefore updated cost function is as given below:

$$cost\ function = \max \sum_{i=1}^{n} f(z) \tag{9}$$

where,

$$Sigmoid\ function\ f(z) = \frac{1}{1+e^{-z}},\quad 0 \le f(z) \ge 1 \tag{10}$$

Here $z = y_i \times w^t x_i$.

Another supervised probabilistic decision-making algorithm (Decision tree) [32] is used for regression and classification of targets based on the training features. The probability tree is penned down using the top-down divide and conquer method. The root of the tree is selected on the variable that provides the maximum information gain, mathematically represented in Eqs. (1), (2) and the next nodes are recursively selected in this manner. The goal is to make the tree as small as possible. Different attribute selection method is applied for selecting the node like GINI Index, shown in Eq. (3). Under Alzheimer's disease detection, pruning is applied at least two instances in leaves and at least five instances in internal nodes with maximum depth 100, also performed splitting and it will stop when majority reaches 95% (in classification only).

Many times, a single tree is not effective to capture mapping to target variables with observations. Therefore, supervised ensemble method such as random forest [33] is also applied over decision trees for decision making in which some number of randomly created decision trees are employed in a small sample of training data and majority decision is used for prediction/classification of target variables. Another ensemble classifier AdaBoost [35] is applied in Alzheimer's disease dataset in which initially some weight is assigned to training sample (Eq. 11) then numerous poorly performing classifiers (base learners: decision tree) are applied sequentially and trains through samples of training data and updates weight according to Eqs. (12) and (13). Afterward, these classifiers combine to get a strong classifier so that high accuracy and less loss are measured.

$$Initial\ Weight\ Estimation(W_I) = \frac{1}{Number\ of\ sample\ (n)} \tag{11}$$

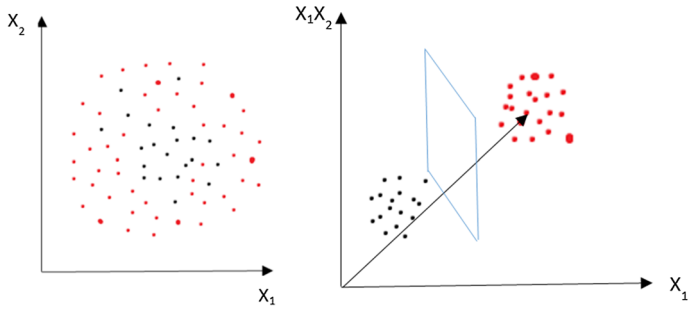

Fig. 5 Non-linear separable data transformed to linearly separable high dimensional space

$$Performance\ of\ stamp\ (p) = \frac{1}{2}\log_e\left(\frac{1 - Total\ Error}{Total\ Error}\right) \tag{12}$$

$$Updated\ Weight = \begin{cases} W_I \times e^{p} & for\ misscalssified\ sample \\ W_I \times e^{-p} & for\ correctly\ calssified\ sample \end{cases} \tag{13}$$

Alzheimer's disease dataset is also analyzed using a support vector machine [16], the idea is to map the feature vectors non-linearly into another space and learn a linear classifier there. The linear classifier in the new space would be an appropriate non-linear classifier in the original space, shown in Fig. 5.

The separating hyperplane given by SVM maximizes the separation between classes [36]. It effectively maps original feature vectors into high dimensional space. Hence, it learns non-linear discriminant functions. Though mapping to high dimensional space requires polynomial computation, therefore SVM uses kernel functions, thus it needs to solve a quadratic optimization problem. In this work, radial basis function kernel has been considered for classification with numerical tolerance of 0.001 and iteration limit of 100, mathematically, it is represented in Eq. (14). The RBF kernel applies at two examples x_1 and x_2, these are represented by way of feature vectors in input hyper-plan, is shown as:

$$f(x_1, x_2) = \exp\left(-\frac{||x_1 - x_2||^2}{2\sigma^2}\right) \tag{14}$$

In the case of linearly separable data, the optimization function is shown in Eq. (15).

$$\left(w^*, b^*\right) = \min \frac{||w||}{2} + c\sum_{i=1}^{n} \varepsilon_i \tag{15}$$

where c = 1.0 and $\varepsilon = 0.1$ represent several errors and the value of error respectively.

For effective classification of Alzheimer's disease, a biologically inspired artificial neural network (ANNs) has been trained under the dataset. It works as the learning process of human brains [20] in which each neuron receives signals, processes them,

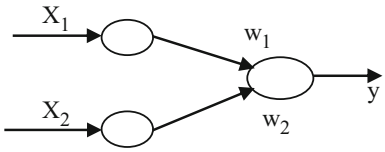

Fig. 6 Simple artificial neuron net

and transmits them to the next nearest neuron and takes decisions. The neural network comprises three layers input layer, hidden layer, and output layer. Several Hidden layers can be inserted according to the optimization level needed. The basic model represented in Fig. 6:

ANN model learns through forward and backward propagation [37]. In forward propagation, the inputs are processed into hidden neurons to get output after applying activation function at each neuron, and passing of data to further neurons. After that, loss is estimated at the output node, as shown in Eq. (16) and if there is an error the signal is propagated backward to each neuron. The goal is to update the weights for minimizing error using the computing gradient of the error function concerning the weight, mathematically represented as chain rule in Eq. (17). Similarly, ANN trains and repeats until convergence.

$$Error\ function\ (E_{total}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \overline{y}_i)^2 \tag{16}$$

Now calculate the gradient of the error function by taking the partial derivative of E_{total} with respect to the concerning weights stated below.

$$\frac{\partial E_{total}}{\partial w_{\mathrm{i}}} = \frac{\partial E_{total}}{\partial \overline{y}} \times \frac{\partial \overline{y}}{\partial net_{input1}} \times \frac{\partial net_{input1}}{\partial w_{\mathrm{i}}} \tag{17}$$

New weights are represented as given below:

$$w_i = w_i - \left(Learning\ rate \times \frac{\partial E_{total}}{\partial w_{\mathrm{i}}}\right)$$

For analysis, ANN has used 100 hidden layers with ReLu activation function and SGD solver. Here learning rate is 0.0001 and the maximum iteration is 300 with replicable training.

Similarly, the Naïve Bayes [34] classification technique is also considered under this work it is based on the probabilistic Bayes theorem and the hypothesis is that all the predictors play an independent and equal contribution to the calculation of the outcomes. Mathematically, Bayes theorem is represented as in Eq. (18),

$$P(y|x_1, x_2, \ldots x_n) = \frac{P(x_1|y)P(x_2|y)\ldots P(x_n|y)P(y)}{P(x_1)P(x_2)\ldots P(x_n)} \tag{18}$$

The above equation shows that, finding out the probability of y (hypothesis), given that $x_1, x_2, \ldots x_n$ (evidence) have already occurred that is called posterior probability

$P(y|x_1)$. $P(y)$ is the prior probability of y, $P(x_1|y)$ is the likelihood here find out the probability of predictor/features given hypothesis and $P(x_1)$ is the prior probability of features.

4 Results and Analysis

In this section, the results of different classification algorithms with parameter confinement as described in previous sections are compared and each classifier is evaluated in terms of ROC-AUC curve, classification accuracy (CA), F1 Score, recall, and precision. Integrated 373 instances are available in the data set with 13 features, with 3 class classification problems. As per the Vapnik-Chervonenkis dimension guidelines that provide a loose bound on several examples, (13X10) 130 instances are sufficient for learning hyper-planes. In addition to that, all the experiments are carried out with tenfold cross-validations, thus making the training set sufficiently large for generalized risk minimization as well as to lower down the true risk on unseen data.

In a fundamental sense, there is no universal model that always performs better than other models for every problem. "No free lunch theorem" formalizes the fact that no learning algorithm is inherently superior. In practice, the performance of the classifier is dependent (unconditionally reliant) on the features used [38]. So, the objective is to find a good subset of features, which is a small subset of features that has the best correlations with the class labels. The results of classifiers are generated by the top 3, 4, 5, 6, and 7 features. The most relevant features are selected by applying feature selection algorithms depicted in Table 2. And the results were obtained as represented in Tables 4, 5, 6 and 7. Tables 2 and 3 classifier outperforms the other classifiers for different numbers feature set and feature selection techniques. The results of best-performing classifiers in the form of accuracy are highlighted in Fig. 7 where classification accuracy is plotted for machine learning techniques for top-rated 4 features, which has been identified after complex feature selection techniques. Here, logistic regression and KNN classifier illustrated higher accuracy (90%). Top-rated four feature maps sufficiently good correlation with the class labels, it is needless to add more features to learn classifiers. Because it adds bias to models which essentially leads to over-fit the model. Formally, it can be stated that the optimum number of features needed to learn all the classifier is four and any more than that leads to adding more computational costs without adding much accuracy.

The interesting finding of this study suggests that Clinical Dementia Rating (CDR) is the most prominent feature to measure the severity of AD. Along with CDR, the Mini-Mental State Examination (MMSE) results helps in the diagnosis of AD more accurately. Socioeconomic status (SES) and normalized whole brain volume (nWBV) in presence of high CDR and MMSE increase the chance of AD. These four attributes are even capable of classifying AD versus non-AD subjects with more than 90% accuracy. The noteworthy contribution of the study is that CDR, MMSE, and SES are interview-based scores, potentially reducing the cost of diagnosis. Using such information in machine learning models helps to identify subjects at higher risk.

There are also techniques to transform the original feature space into a new feature space. It certainly helps to improve the features. For example, to make them

Table 4 Classification results for top-rated 3, 4, 5, 6 and 7 features using information gain

Model	Top-rated 3 features					Top-rated 4 features					Top-rated 5 features				
	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall
kNN	0.902	0.818	0.804	0.794	0.82	0.914	0.847	0.835	0.826	0.847	0.898	0.799	0.785	0.784	0.8
Decision tree	0.921	0.887	0.869	0.861	0.89	0.896	0.853	0.855	0.858	0.853	0.881	0.855	0.855	0.855	0.86
SVM	0.907	0.89	0.844	0.803	0.89	0.928	0.885	0.841	0.802	0.885	0.947	0.887	0.847	0.837	0.89
SGD	0.88	0.895	0.849	0.807	0.9	0.88	0.895	0.849	0.807	0.895	0.88	0.895	0.849	0.807	0.9
Random forest	0.921	0.885	0.865	0.855	0.89	0.926	0.879	0.865	0.856	0.879	0.939	0.893	0.873	0.867	0.89
Neural network	0.937	0.893	0.856	0.858	0.89	0.937	0.887	0.854	0.855	0.887	0.946	0.893	0.856	0.858	0.89
Naive Bayes	0.939	0.887	0.874	0.867	0.89	0.938	0.879	0.863	0.854	0.879	0.94	0.882	0.868	0.859	0.88
Logistic regression	0.942	0.898	0.868	0.871	0.9	0.939	0.901	0.869	0.881	0.901	0.943	0.898	0.864	0.872	0.9
AdaBoost	0.924	0.871	0.856	0.844	0.87	0.844	0.823	0.82	0.817	0.823	0.845	0.815	0.821	0.828	0.82

Model	Top-rated 6 features					Top-rated 7 features				
	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall
kNN	0.744	0.633	0.611	0.599	0.633	0.729	0.638	0.618	0.607	0.638
Decision tree	0.904	0.855	0.855	0.855	0.855	0.895	0.863	0.862	0.862	0.863
SVM	0.941	0.895	0.874	0.874	0.895	0.96	0.895	0.868	0.874	0.895
SGD	0.883	0.898	0.855	0.908	0.898	0.889	0.903	0.867	0.913	0.903

Table 4 (continued)

Model	Top-rated 6 features					Top-rated 7 features				
	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall
Random forest	0.944	0.901	0.881	0.88	0.901	0.946	0.885	0.861	0.851	0.885
Neural network	0.949	0.887	0.861	0.862	0.887	0.953	0.903	0.875	0.888	0.903
Naive Bayes	0.941	0.893	0.882	0.876	0.893	0.944	0.882	0.876	0.871	0.882
Logistic regression	0.934	0.887	0.862	0.871	0.887	0.924	0.826	0.798	0.789	0.826
AdaBoost	0.872	0.853	0.853	0.853	0.853	0.868	0.847	0.848	0.849	0.847

Table 5 Classification results for top- rated 3, 4, 5, 6 and 7 features using information gain ratio

Model	Top-rated 3 features					Top-rated 4 features					Top-rated 5 features				
	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall
kNN	0.895	0.861	0.828	0.797	0.861	0.914	0.847	0.835	0.826	0.847	0.875	0.799	0.777	0.767	0.799
Decision tree	0.898	0.882	0.842	0.806	0.882	0.896	0.853	0.855	0.858	0.853	0.898	0.855	0.855	0.855	0.855
SVM	0.893	0.895	0.849	0.807	0.895	0.928	0.885	0.841	0.802	0.885	0.951	0.890	0.857	0.857	0.890
SGD	0.880	0.895	0.849	0.807	0.895	0.880	0.895	0.849	0.807	0.895	0.880	0.895	0.849	0.807	0.895
Random forest	0.900	0.885	0.843	0.806	0.885	0.926	0.879	0.865	0.856	0.879	0.936	0.882	0.857	0.844	0.882
Neural network	0.919	0.895	0.849	0.807	0.895	0.937	0.887	0.854	0.855	0.887	0.943	0.893	0.856	0.858	0.893
Naive Bayes	0.918	0.893	0.860	0.855	0.893	0.938	0.879	0.863	0.854	0.879	0.940	0.890	0.871	0.863	0.890
Logistic regression	0.917	0.890	0.851	0.833	0.890	0.939	0.901	0.869	0.881	0.901	0.941	0.895	0.866	0.864	0.895
AdaBoost	0.887	0.718	0.757	0.826	0.718	0.844	0.823	0.820	0.817	0.823	0.862	0.842	0.841	0.840	0.842

Model	Top-rated 6 features					Top-rated 7 features				
	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall
kNN	0.899	0.820	0.807	0.807	0.820	0.729	0.638	0.618	0.607	0.638
Decision tree	0.881	0.855	0.852	0.848	0.855	0.896	0.871	0.868	0.865	0.871
SVM	0.959	0.893	0.859	0.868	0.893	0.959	0.895	0.868	0.874	0.895
SGD	0.880	0.895	0.849	0.807	0.895	0.889	0.903	0.867	0.913	0.903

Table 5 (continued)

Model	Top-rated 6 features					Top-rated 7 features				
	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall
Random forest	0.938	0.885	0.861	0.852	0.885	0.954	0.901	0.883	0.882	0.901
Neural network	0.951	0.890	0.853	0.855	0.890	0.953	0.903	0.871	0.895	0.903
Naive Bayes	0.941	0.887	0.878	0.872	0.887	0.944	0.882	0.876	0.871	0.882
Logistic regression	0.946	0.901	0.869	0.881	0.901	0.924	0.828	0.800	0.793	0.828
AdaBoost	0.863	0.842	0.843	0.843	0.842	0.865	0.845	0.845	0.846	0.845

Table 6 Classification results for top- rated 3, 4, 5, 6 and 7 features using Gini Index

Model	Top-rated 3 features					Top-rated 4 features					Top-rated 5 features				
	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall
kNN	0.906	0.877	0.846	0.827	0.877	0.892	0.839	0.825	0.818	0.839	0.898	0.799	0.785	0.784	0.799
Decision tree	0.882	0.831	0.826	0.820	0.831	0.888	0.850	0.848	0.846	0.850	0.881	0.853	0.853	0.854	0.853
SVM	0.894	0.895	0.849	0.807	0.895	0.902	0.895	0.849	0.807	0.895	0.943	0.887	0.847	0.837	0.887
SGD	0.880	0.895	0.849	0.807	0.895	0.880	0.895	0.849	0.807	0.895	0.880	0.895	0.849	0.807	0.895
Random forest	0.909	0.866	0.854	0.845	0.866	0.899	0.890	0.871	0.864	0.890	0.930	0.874	0.860	0.850	0.874
Neural network	0.929	0.895	0.849	0.807	0.895	0.929	0.895	0.849	0.807	0.895	0.942	0.890	0.850	0.839	0.890
Naive Bayes	0.922	0.877	0.843	0.818	0.877	0.926	0.895	0.872	0.868	0.895	0.940	0.882	0.868	0.859	0.882
Logistic regression	0.915	0.895	0.849	0.807	0.895	0.918	0.890	0.846	0.806	0.890	0.943	0.898	0.864	0.872	0.898
AdaBoost	0.825	0.791	0.800	0.811	0.791	0.847	0.818	0.824	0.830	0.818	0.845	0.815	0.821	0.828	0.815

Model	Top-rated 6 features					Top-rated 7 features				
	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall
kNN	0.899	0.820	0.807	0.807	0.820	0.729	0.638	0.618	0.607	0.638
Decision tree	0.881	0.853	0.850	0.847	0.853	0.895	0.866	0.864	0.863	0.866
SVM	0.960	0.893	0.859	0.868	0.893	0.960	0.895	0.868	0.874	0.895
SGD	0.880	0.895	0.849	0.807	0.895	0.889	0.903	0.867	0.913	0.903

Table 6 (continued)

Model	Top-rated 6 features					Top-rated 7 features				
	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall
Random forest	0.934	0.885	0.865	0.855	0.885	0.945	0.882	0.859	0.848	0.882
Neural network	0.947	0.893	0.856	0.858	0.893	0.956	0.898	0.864	0.872	0.898
Naive Bayes	0.941	0.887	0.878	0.872	0.887	0.944	0.882	0.876	0.871	0.882
Logistic regression	0.946	0.901	0.869	0.881	0.901	0.923	0.826	0.798	0.789	0.826
AdaBoost	0.862	0.839	0.841	0.842	0.839	0.872	0.855	0.853	0.851	0.855

Table 7 Classification results for top- rated 3, 4, 5, 6 and 7 features using Chi-squared

Model	Top-rated 3 features					Top-rated 4 features					Top-rated 5 features				
	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall
kNN	0.898	0.799	0.785	0.784	0.799	0.892	0.839	0.825	0.818	0.839	0.898	0.799	0.785	0.784	0.799
Decision tree	0.881	0.853	0.853	0.854	0.853	0.888	0.850	0.848	0.846	0.850	0.881	0.853	0.853	0.854	0.853
SVM	0.946	0.887	0.847	0.837	0.887	0.903	0.895	0.849	0.807	0.895	0.948	0.887	0.847	0.837	0.887
SGD	0.880	0.895	0.849	0.807	0.895	0.880	0.895	0.849	0.807	0.895	0.880	0.895	0.849	0.807	0.895
Random forest	0.929	0.887	0.867	0.858	0.887	0.916	0.901	0.879	0.879	0.901	0.928	0.882	0.860	0.849	0.882
Neural network	0.942	0.890	0.850	0.839	0.890	0.929	0.895	0.849	0.807	0.895	0.942	0.890	0.850	0.839	0.890
Naive Bayes	0.940	0.882	0.868	0.859	0.882	0.926	0.895	0.872	0.868	0.895	0.940	0.882	0.868	0.859	0.882
Logistic regression	0.943	0.898	0.864	0.872	0.898	0.918	0.890	0.846	0.806	0.890	0.943	0.898	0.864	0.872	0.898
AdaBoost	0.845	0.815	0.821	0.828	0.815	0.847	0.818	0.824	0.830	0.818	0.845	0.815	0.821	0.828	0.815

Model	Top-rated 6 features					Top-rated 7 features				
	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall
kNN	0.744	0.633	0.611	0.599	0.633	0.729	0.638	0.618	0.607	0.638
Decision tree	0.904	0.855	0.855	0.855	0.855	0.895	0.863	0.862	0.862	0.863
SVM	0.943	0.895	0.874	0.874	0.895	0.960	0.895	0.868	0.874	0.895
SGD	0.889	0.903	0.867	0.913	0.903	0.889	0.903	0.867	0.913	0.903

Table 7 (continued)

Model	Top-rated 6 features					Top-rated 7 features				
	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall
Random forest	0.942	0.906	0.888	0.889	0.906	0.952	0.887	0.863	0.855	0.887
Neural network	0.950	0.890	0.856	0.865	0.890	0.953	0.901	0.873	0.879	0.901
Naive Bayes	0.941	0.893	0.882	0.876	0.893	0.944	0.882	0.876	0.871	0.882
Logistic regression	0.934	0.887	0.862	0.871	0.887	0.924	0.826	0.798	0.789	0.826
AdaBoost	0.877	0.858	0.859	0.861	0.858	0.862	0.839	0.841	0.843	0.839

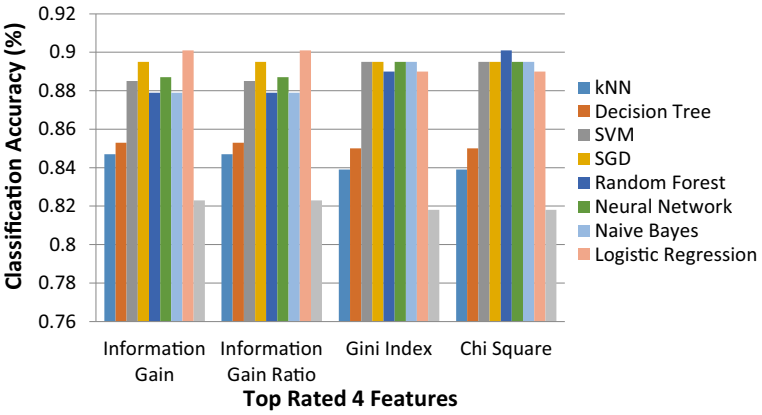

Fig. 7 Accuracy versus top rated 4 features

uncorrelated or reduce the dimensionality of the feature vector without losing too much information. One such technique is PCA in which all the features are linearly transformed and projected to lower-dimensional space. It essentially maximizes the variance of projected data. The results of different principal components with the percentage of variability covered by new projected data are depicted in Table 8 and Fig. 8. Again, looking at these tables, covering only 59.3% variability, the maximum performance of the classifier is achieved. So, it can be stated that the top four principal components are again sufficient to map the correlation with class labels.

5 Conclusion

The machine learning algorithms used in this paper are standards and successfully applied in classification problems. Along with classification algorithms, different feature selection and dimension reduction techniques are used for diffing out more relevant features than others for decision making and thus reducing the training time of the classification algorithms. The contribution of this study is helpful in the exploration of different machine learning algorithms in the field of applied research in brain health maintenance as well as disease prevention using an MRI dataset. This paper also suggests the optimal parameters to AD investigators i.e. the symptoms that contribute well in decision making for machine learning algorithms. Accordingly, a lower number of features has been identified which essentially lowers the cost and time for disease diagnosis. In a nutshell the effectiveness of the prediction of the classifiers depends on the number of features and feature selection techniques. In this study, Top-rated four features namely CDR, SES, nWBV, and EDUC are identified for decision making for AD that map sufficiently accurate correlation with the class labels and an approximately 90% accuracy. Finally, combining the critical feature selection techniques with the appropriate classifier yields better results. Furthermore, the work can be extended

Table 8 Classification results for top- rated 3, 4, 5, 6 and 7 principal components with variation of 49.5%, 59.3%, 67.6%, 75% and 82.3% respectively

Model	Top-rated 3 features (variation of 49.5%)					Top-rated 4 features (variation of 59.3%)					Top-rated 5 features (variation of 67.6%)				
	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall
kNN	0.546	0.477	0.467	0.459	0.477	0.546	0.477	0.467	0.459	0.477	0.546	0.477	0.467	0.459	0.477
Decision tree	0.936	0.882	0.872	0.865	0.882	0.936	0.882	0.872	0.865	0.882	0.936	0.882	0.872	0.865	0.882
SVM	0.964	0.898	0.876	0.879	0.898	0.964	0.898	0.876	0.879	0.898	0.963	0.898	0.876	0.879	0.898
SGD	0.892	0.906	0.873	0.915	0.906	0.892	0.906	0.873	0.915	0.906	0.892	0.906	0.873	0.915	0.906
Random forest	0.952	0.895	0.871	0.874	0.895	0.953	0.901	0.876	0.878	0.901	0.948	0.901	0.878	0.886	0.901
Neural network	0.960	0.906	0.877	0.900	0.906	0.960	0.906	0.877	0.900	0.906	0.960	0.906	0.877	0.900	0.906
Naive Bayes	0.941	0.853	0.858	0.864	0.853	0.941	0.853	0.858	0.864	0.853	0.941	0.853	0.858	0.864	0.853
Logistic regression	0.794	0.678	0.640	0.624	0.678	0.794	0.678	0.640	0.624	0.678	0.794	0.678	0.640	0.624	0.678
AdaBoost	0.872	0.847	0.851	0.856	0.847	0.872	0.847	0.851	0.856	0.847	0.872	0.847	0.851	0.856	0.847

Model	Top-rated 6 features (variation of 75%)					Top-rated 7 features (variation of 82.3%)				
	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall
kNN	0.546	0.477	0.467	0.459	0.477	0.546	0.477	0.467	0.459	0.477
Decision tree	0.936	0.882	0.872	0.865	0.882	0.936	0.882	0.872	0.865	0.882
SVM	0.965	0.898	0.876	0.879	0.898	0.965	0.898	0.876	0.879	0.898
SGD	0.892	0.906	0.873	0.915	0.906	0.892	0.906	0.873	0.915	0.906

Table 8 (continued)

Model	Top-rated 6 features (variation of 75%)					Top-rated 7 features (variation of 82.3%)				
	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall
Random forest	0.943	0.895	0.872	0.868	0.895	0.955	0.903	0.880	0.894	0.903
Neural network	0.960	0.906	0.877	0.900	0.906	0.960	0.906	0.877	0.900	0.906
Naive Bayes	0.941	0.853	0.858	0.864	0.853	0.941	0.853	0.858	0.864	0.853
Logistic regression	0.794	0.678	0.640	0.624	0.678	0.794	0.678	0.640	0.624	0.678
AdaBoost	0.872	0.847	0.851	0.856	0.847	0.872	0.847	0.851	0.856	0.847

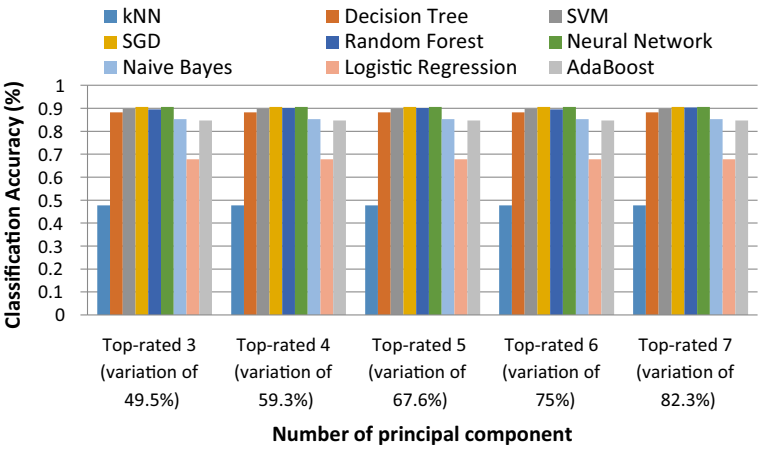

Fig. 8 Classification accuracy versus number of principal component

by applying other ensemble classification techniques on selected features to increase the accuracy of the diagnosis.

Acknowledgements Data were provided by OASIS Longitudinal: Principal Investigators: D. Marcus, R, Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382.

Author Contributions All authors have been personally and actively involved in substantial work leading to the paper and will take public responsibility for its content. Siddhartha Kumar Arjaria and Abhishek Singh Rathore defined methodology and performed result analysis. Dhananjay Bisen worked on related work section. Sanjib Bhattacharya identified the problem and verified results.

Funding The Authors confirmed that they received no funding from any institution or any government.

Data availability All the data and code are available on github: https://github.com/abhishekatujjain/Alzheimer/

Code availability All the data and code are available on github: https://github.com/abhishekatujjain/Alzheimer/

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical Statement The Authors consciously assure that for the manuscript is the authors' own original work, which has not been previously published elsewhere. The paper is not currently being considered for publication elsewhere. The paper reflects the authors' own research and analysis in a truthful and complete manner. The paper properly credits the meaningful contributions of co-authors and co-researchers. The results are appropriately placed in the context of prior and existing research. All sources used are properly disclosed (correct citation). Literally copying of text must be indicated as such by using quotation marks and giving proper reference. The violation of the Ethical Statement rules may result in severe consequences.

References

1. Saravanakumar S, Thangaraj P (2019) A voxel based morphometry approach for identifying Alzheimer from MRI images. Cluster Comput 22:14081–14089. https://doi.org/10.1007/s10586-018-2236-6
2. Gaugler J, James B, Marin A (2019) 2019 Alzheimer's disease facts and figures
3. Ledig C, Schuh A, Guerrero R et al (2018) Structural brain imaging in Alzheimer's disease and mild cognitive impairment: biomarker analysis and shared morphometry database. Sci Rep 8:11258. https://doi.org/10.1038/s41598-018-29295-9
4. Xu N, Shen Y, Zhu YY et al (2017) Internet of things, real-time decision making, and artificial intelligence. In: Mishra D, Buyya R, Mohapatra P, Patnaik S (eds) Medical image computing and computer assisted intervention-MICCAI 2017. Springer, Cham, pp 107–115
5. Tien JM (2017) Internet of things, real-time decision making, and artificial intelligence. Ann Data Sci 4:149–178. https://doi.org/10.1007/s40745-017-0112-5
6. Zhang M, Yang Y, Zhang H et al (2016) L2, p-norm and sample constraint based feature selection and classification for AD diagnosis. Neurocomputing 195:104–111. https://doi.org/10.1016/j.neucom.2015.08.111
7. Zhang M, Yang Y, Shen F et al (2017) Multi-view feature selection and classification for Alzheimer's disease diagnosis. Multimed Tools Appl 76:10761–10775. https://doi.org/10.1007/s11042-015-3173-5
8. Luo P, Kang G, Xu X (2020) A novel feature selection and classification method of Alzheimer's disease based on multi-features in MRI. In: Proceedings of the 2020 10th international conference on bioscience, biochemistry and bioinformatics. Association for Computing Machinery, New York, NY, USA, pp 114–119
9. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–1238. https://doi.org/10.1109/TPAMI.2005.159
10. Zheng X, Shi J, Zhang Q et al (2017) Improving MRI-based diagnosis of Alzheimer's disease via an ensemble privileged information learning algorithm. In: Proceedings of 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017), pp 456–459
11. Ji H, Liu Z, Yan WQ, Klette R (2019) Early diagnosis of Alzheimer's disease using deep learning. In: Proceedings of the 2nd international conference on control and computer vision. Association for Computing Machinery, New York, NY, USA, pp 87–91
12. Valliani A, Soni A (2017) Deep residual nets for improved Alzheimer's diagnosis. In: Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics. Association for Computing Machinery, New York, NY, USA, p 615
13. Cherdal S, Mouline S (2016) Petri nets for modelling and analysing a complex system related to Alzheimer's disease. In: Proceedings of the 31st annual ACM symposium on applied computing. Association for Computing Machinery, New York, NY, USA, pp 309–312
14. McCrackin L (2018) Early detection of Alzheimer's disease using deep learning. In: Bagheri E, Cheung JCK (eds) Advances in artificial intelligence. Springer, Cham, pp 355–359
15. Liu M, Zhang J, Adeli E, Shen D (2017) Deep multi-task multi-channel learning for joint classification and regression of brain status. In: International conference on medical image computing and computer-assisted intervention-MICCAI, vol 10435, pp 3–11. https://doi.org/10.1007/978-3-319-66179-7_1
16. Zhao Y, He L (2015) Deep Learning in the EEG diagnosis of Alzheimer's disease. In: Jawahar CV, Shan S (eds) Computer vision-ACCV 2014 workshops. Springer, Cham, pp 340–353
17. Sun X, Hu L, Yao Y, Wang Y (2017) GSplit LBI: taming the procedural bias in neuroimaging for disease prediction. In: Descoteaux M, Maier-Hein L, Franz A et al (eds) Medical image computing and computer assisted intervention-MICCAI 2017. Springer, Cham, pp 107–115
18. Cao P, Liu X, Yang J et al (2017) Sparse multi-kernel based multi-task learning for joint prediction of clinical scores and biomarker identification in Alzheimer's disease. In: Descoteaux M, Maier-Hein L, Franz A et al (eds) Medical image computing and computer assisted intervention-MICCAI 2017. Springer, Cham, pp 195–202
19. Liu X, Cao P, Gonçalves AR et al (2018) Modeling Alzheimer's disease progression with fused Laplacian sparse group lasso. ACM Trans Knowl Discov Data. https://doi.org/10.1145/3230668
20. Xu N, Shen Y, Zhu Y (2019) A multi-task learning framework for automatic early detection of Alzheimer's. In: Li G, Yang J, Gama J et al (eds) Database systems for advanced applications. Springer, Cham, pp 240–243

21. Zhang P, Shi B, Smith CD, Liu J (2017) Nonlinear feature space transformation to improve the prediction of MCI to AD conversion. In: Descoteaux M, Maier-Hein L, Franz A et al (eds) Medical image computing and computer assisted intervention-MICCAI 2017. Springer, Cham, pp 12–20
22. Zhu X, Thung K-H, Adeli E et al (2017) Maximum mean discrepancy based multiple kernel learning for incomplete multimodality neuroimaging data. In: Descoteaux M, Maier-Hein L, Franz A et al (eds) Medical image computing and computer assisted intervention-MICCAI 2017. Springer, Cham, pp 72–80
23. Zhu Y, Kim M, Zhu X et al (2017) Personalized diagnosis for Alzheimer's disease. In: Descoteaux M, Maier-Hein L, Franz A et al (eds) Medical image computing and computer assisted intervention-MICCAI 2017. Springer, Cham, pp 205–213
24. Gamberger D, Ženko B, Mitelpunkt A et al (2016) Clusters of male and female Alzheimer's disease patients in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Brain Inform 3:169–179. https://doi.org/10.1007/s40708-016-0035-5
25. Liu S (2017) Alzheimer's disease staging and prediction. In: Multimodal neuroimaging computing for the characterization of neurodegenerative disorders. Springer, Singapore, pp 95–108
26. Abásolo D, Hornero R, Espino P et al (2006) Entropy analysis of the EEG background activity in Alzheimer's disease patients. Physiol Meas 27:241–253. https://doi.org/10.1088/0967-3334/27/3/003
27. Vecchio F, Miraglia F, Piludu F et al (2017) "Small World" architecture in brain connectivity and hippocampal volume in Alzheimer's disease: a study via graph theory from EEG data. Brain Imaging Behav 11:473–485. https://doi.org/10.1007/s11682-016-9528-3
28. Kawahara J, Brown CJ, Miller SP et al (2017) BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment. Neuroimage 146:1038–1049. https://doi.org/10.1016/j.neuroimage.2016.09.046
29. Palafox GDL, Ortíz ALS, Melendez OM, et al (2017) Hippocampal segmentation using mean shift algorithm. In: Proceeding of SPIE
30. Chen X, Zhao D, Zhong W (2019) Auxiliary recognition of Alzheimer's disease based on Gaussian probability brain image segmentation model. In: Ning H (ed) Cyberspace data and intelligence, and cyber-living, syndrome, and health. Springer, Singapore, pp 513–520
31. Marcus DS, Fotenos AF, Csernansky JG et al (2010) Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. J Cogn Neurosci 22:2677–2684. https://doi.org/10.1162/jocn.2009.21407
32. Brownlee J (2019) Information gain and mutual information for machine learning. In: Machine learning mastery. https://machinelearningmastery.com/information-gain-and-mutual-information/. Accessed 3 March 2021
33. Brown SD, Myles AJ (2009) Decision tree modeling
34. Dash SS, Nayak SK, Mishra D (2021) A review on machine learning algorithms. In: Mishra D, Buyya R, Mohapatra P, Patnaik S (eds) Intelligent and cloud computing. Springer, Singapore, pp 495–507
35. Chaubey G, Bisen D, Arjaria S, Yadav V (2021) Thyroid disease prediction using machine learning approaches. Natl Acad Sci Lett 44:233–238. https://doi.org/10.1007/s40009-020-00979-z
36. Shi Y, Tian Y, Kou G, et al (2011) Support vector machines for classification problems. In: Optimization based data mining: theory and applications. Springer, London, pp 3–13
37. Olson DL, Shi Y (2007) Introduction to business data mining. McGraw-Hill/Irwin
38. Shi Y (2022) Feature selection. In: Advances in big data analytics: theory, algorithms and practices. Springer, Singapore, pp 249–304