EasyChair Preprint

№ 9756

Analysis of Machine Learning Algorithms for Alzheimer's Disease Classification and Prediction

Dhwani Modi and Pushpendra Singh Sisodia

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

February 21, 2023

Analysis Of Machine Learning Algorithms For Alzheimer's Disease Classification And Prediction

Dhwani Modi[1, a)], Pushpendra Singh Sisodia[2, b)]

Indus Institute of Technology & Engineering,
Indus University
Ahmedabad, Gujarat, India[1,2]

a) dhwanimodi56@gmail.com

b) pushpendrasingh.ce@indusuni.ac.in

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder in which dementia and cognitive decline gradually. AD is the major disease that is responsible for memory loss and cognitive impairment in older people. Till now there no known cure is available for AIAD. Early detection of AD and proper time treatment may prevent AD. Early detection of AD is possible using Machine Learning (ML) and MRI data. In this study, supervised ML classification along with longitudinal brain MRI data are used to predict and classify AD. The six states of art supervised classifiers named Random Forest (RF), XGBoost, Support Vector Machine (SVM), ExtraTree Classifier, Naive Bayes (NB) Classifier, and KNearest Neighbours (KNN) are used to classify and predict. The performance of the ML classifier was evaluated using performance metrics such as precision, recall, accuracy, f1-score and ROC/AUC curve. Among all six ML classifiers, the ExtraTree performed the best with the highest classification accuracy of 87%.

Keywords: Alzheimer's disease (AD), Dementia, Machine Learning, Classification, Prediction, Performance

1.INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that slows down cognition. Cognitive slowdown results in dementia, abnormal behaviour, and personality changes, and sometimes lead to death [1][8]. According to the World Alzheimer Report-2019, more than 57 million people are suffering from dementia worldwide. It is estimated that there will be 131.5 million elderly people suffering from dementia by 2050 globally [2]. Globally, dementia is the seventh leading cause of death among all diseases [3]. In developing nations, the frequency of AD is around 5% among the age of 65 and 30% among people over 85. AD patients typically live around 8 and 10 years after diagnosis, while the illness can extend up to 20 years [4]. Understanding the disease's course over a long period is one of the main issues in Alzheimer's disease research, and it is ideal to identify the trend before the emergence of clinical signs. Mild cognitive impairment (MCI) is a stage between normal cognitive function and AD. The MRI data and Machine Learning classification are helpful to identify the stage between normal cognitive function and AD. A recent study has used several techniques of machine learning such as Decision Tree, Random Forest, Support Vector Machine, Gradient Boosting, and voting classifiers to extract the best features for Alzheimer's disease prediction using a longitudinal dataset [5]. Some studies used deep learning and transfer learning models to classify and predict early AD symptoms [8-10]. The proposed work presented promising results with the best validation accuracy of 83%. Another study proposed a framework based on supervised learning classifiers in the dementia subject categorization in AD or non-AD based on longitudinal brain MRI features. The result shows that the gradient boosting algorithm outperforms the highest accuracy of 97.58% [6]. Some study assesses the impact of a dimensionality reduction technique involving Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Neighbourhood Component Analysis (NCA), Factor Analysis (FA), Fast Independent Component Analysis (FastICA) applied to the cross-sectional MRI data. The study attained an overall accuracy of 87% [7]. In this study, longitudinal MRI data such as age, gender, education, socioeconomic status (SES), Mini-Mental State Examination (MMSE), estimated total intracranial volume (eTIV), Clinical Dementia Rating (CDR), Atlas Scaling Factor (ASF) and normalized whole brain volume (nWBV) is used. The six states of art supervised classifiers named Random Forest (RF), XGBoost, Support Vector Machine (SVM), ExtraTree Classifier, Naive Bayes (NB) Classifier, and KNearest Neighbours (KNN) are used to classify and predict. The performance of the ML classifier was evaluated using performance metrics such as precision, recall, accuracy, f1-score and ROC/AUC curve. Among all six ML classifiers, the ExtraTree performed the best with the highest classification accuracy of 87%.

2. DATA AND METHODOLOGY

The Open Access Series of Brain Imaging (OASIS-2) dataset was used. We used exploratory data analysis (EDA) to determine how distinct feature sets correlate with one another. The OASIS-2 dataset includes 373 observations and 15 attributes and is listed in Table 1. In the dataset, the patients are labelled as having dementia or no dementia.

TABLE 1 Dataset Attribute description

Attribute Name	Attribute Description
Subject ID	Subject identification number
MRI ID	Image identification number of an individual subject
Group	Demented/Non-demented/Converted
Visit	Number of subjects visit
MR Delay	Magnetic resonance delay is the period of time before the acquisition of an image
Gender	Male/Female
Hand	Right/Left-Handed
Age	Subject age while scanning

EDUC	Subject educational level
SES	Socioeconomic status
MMSE	Mini-Mental state examination scores
CDR	Clinical dementia rating score
eTIV	Estimated total intracranial volume
nWBV	Normalized whole brain volume output
ASF	Atlas scaling factor

2.1 Data Pre-Processing

The raw MRI data is pre-processed as null value replacement, duplicate data removal and performed data normalization.

2.2 Data Splitting (Train and Test)

The dataset is partitioned into training and testing. The ratio of training and testing of the dataset is 80:20, respectively. 80% of the dataset is used to train the model and 20% of the data is utilized for model validation and prediction.

2.3 Models Used

In this study, the six states of art supervised classifiers named Random Forest (RF), XGBoost, Support Vector Machine (SVM), ExtraTree Classifier, Naive Bayes (NB) Classifier, and KNearest Neighbours (KNN) are used to classify and prediction. A brief explanation of the ML classifiers is discussed in the next section.

2.3.1 Random Forest Classifier (RF)

With the use of the Random Forest method, for each sample taken in the training dataset, a decision tree is formed and then trained using 5-fold cross-validation with different parameter combinations. we use the Gini index to determine how nodes on a decision tree branch should be ordered based on categorization data.

$$\textbf{Gini} = 1 - \sum_{i=1}^{c} (pi)^2 \quad \text{(i)}$$

Here, the number of classes is represented by c, and the relative frequency of the class you are observing in the dataset is represented by pi.

2.3.2 Support Vector Machine (SVM)

The objective of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes. The term "hyperplane" refers to this optimal decision boundary. This model is implemented initially by taking the regularization parameter C = 0.1, and linear basis function as the kernel followed by 5-fold cross-validation. Finally, its classification or prediction performance is studied with the help of a confusion matrix.

2.3.3 ExtraTree Classifier

An extremely Randomized Trees Classifier (Extra Trees Classifier) is a type of ensemble learning technique which fundamentally based on a decision tree. To get its classification result, it aggregates the findings of several de-correlated decision trees gathered in a "forest.". Decision criteria are used for information Gain. First, Entropy will be calculated.

$$\textbf{Entropy (S)} = \sum_{i=1}^{c} - \textbf{pilog}_2(\textbf{p}_i) \quad \text{(ii)}$$

2.3.4 GaussianNB Classifier

The GNB classifier uses the Bayes theorem and is implemented using a mutually independent variable. It is used as a selective classifier for dementia. After the GNB model had been trained, 5-fold cross-validation was performed. The mean and variance of X are necessary to calculate this formula.

$$\textbf{P(X|Y = c)} = \frac{1}{\sqrt{2\pi\sigma 2}} e^{\frac{-(x-\mu)2}{2\sigma 2}} \quad \text{(iii)}$$

Sigma and mu are the continuous variable X's variance and mean as calculated for the specified class c of Y in the formulas above.

2.3.5 XGBoost

XGBoost which stands for Extreme Gradient Boosting is an ensemble learning method. It is a scalable, distributed gradient-boosted decision tree (GBDT). It provides a parallel tree boosting and is the leading machine-learning library for regression, classification, and ranking problems.

The objective function for the XGBoost is given by the following formula.

$$\textbf{obj}(\theta) = \sum_{i=1}^{n} \textbf{l}(\textbf{yi}, \hat{\textbf{y}}\textbf{i}) + \sum_{K=1}^{K} \Omega(fk) \quad \text{(vi)}$$

Where k is the number of trees, f is the functional space of f, and f is the set of possible CARTs. Here, the first term is the loss function and the second is the regularization parameter.

2.3.6 KNeighbours Classifier

The most important challenge is choosing the value of K in the K-nearest neighbour algorithm. A low number of K indicates that noise will have a greater impact on the outcome, which increases the likelihood of overfitting. K = n ^ (1/2) is an easy formula to use when choosing k. We will use the Euclidean metric for measuring distance.

$$\mathbf{d(x, x')} = \sqrt{(x1 - x1')^2 + \ldots + (xn - xn')^2} \quad \text{(v)}$$

3. RESULTS

The study evaluated ML models' performance by calculating Precision, Recall, F1-score, and Accuracy for train data. We have also plotted the AUC-ROC curve for each model.

3.1 Performance Metrics

Precision: It measures how often the prediction is correct when predicting positive instances. Here, TP represents True positive samples and FP represents false positive samples

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

Recall: It is also known as Sensitivity or True Positive Rate. It computes predicted positive classes in the total positive occurrences in the dataset. Here, FN represents false negative samples

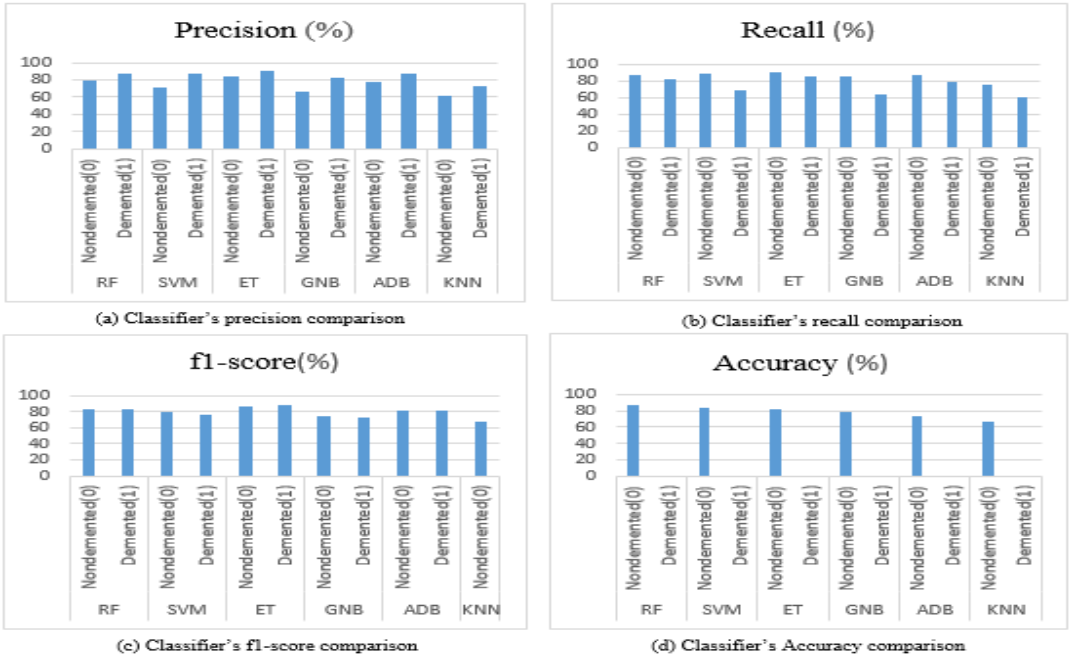
$$\mathbf{Recall} = \frac{TP}{TP+FN} \quad \text{(vii)}$$

F1-Score: It is also known as harmonic mean. It calculates the accuracy of the testing process. The average is calculated using precisions and recall sets. The range of the f1-score lies between 0 to 1.

$$\mathbf{F1\text{-}score} = 2 * \frac{Precision * Recall}{Precision + Recall} \quad \text{(viii)}$$

Accuracy: The simplest measure of performance is classification accuracy. It is defined as the percentage of properly predicted batches and is determined using the formula below.

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



(a) Classifier's precision comparison

(b) Classifier's recall comparison

(c) Classifier's f1-score comparison

(d) Classifier's Accuracy comparison

FIGURE 1. Comparison of performance metrics such as (a) precision, (b) recall, (c) f1-score and (d) accuracy with classification models

3.2 Confusion Matrix

Typically, a confusion matrix is an N × N matrix that is made up of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) samples that are received while classification. The confusion matrices mentioned in figure 2 are recorded during the classification process.

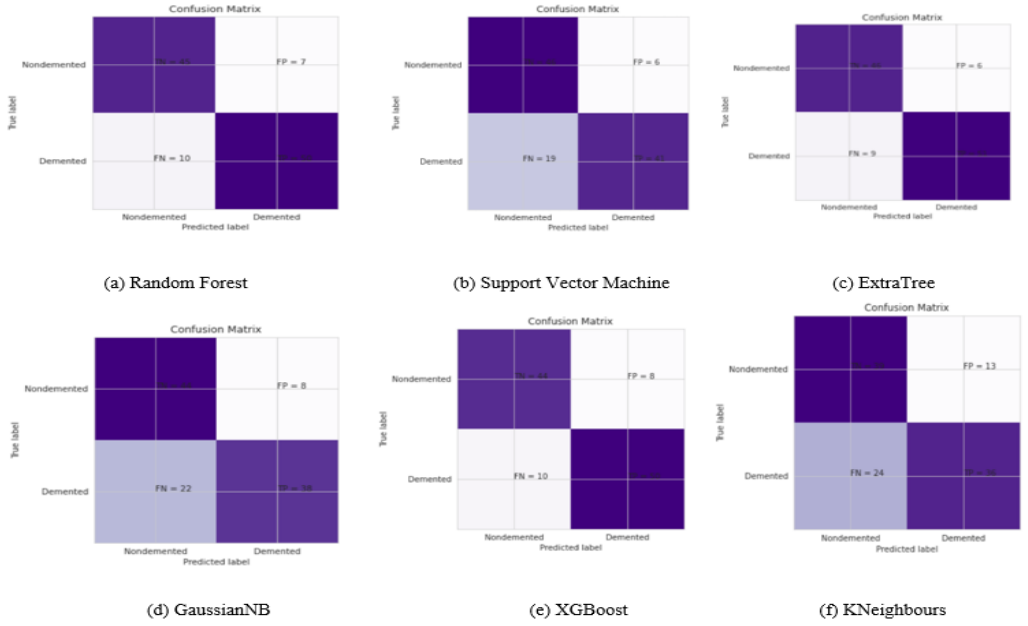


FIGURE 2. The confusion matrix outcome of (a) Random Forest (a) Support Vector Machine (b) ExtraTree (c) GaussianNB(d) XGBoost (e) KNeighbour.

3.3 Performance Metrics

The study included various performance metrics like accuracy, precision, recall and F1 score. To determine the best parameters for each model, 5-fold cross-validation is used. The classifiers were evaluated to determine how accurately they predicted the AD subject's status on a test data subset. All given classifiers perform well in the classification of AD subjects, but ExtraTree outperforms all the classifiers.

TABLE 2. Performance Comparison of different ML classifiers.

Sr. No.	ML Models	Model	Classification	Precision (%)	Recall (%)	f1-score (%)	Support (%)	Accuracy (%)
1.	ExtraTree Classifier	ET	Non-demented(0)	84	90	87	52	87
			Demented(1)	91	85	88	60	
2.	Random Forest Classifier (RF)	RF	Non-demented(0)	80	87	83	52	83
			Demented(1)	88	82	84	60	
3.	AdaBoost classifier	ADB	Non-demented(0)	78	87	82	52	82
			Demented(1)	87	78	82	60	
4.	Support Vector Machine (SVM)	SVM	Non-demented(0)	71	88	79	52	78
			Demented(1)	87	68	77	60	
5.	GaussianNB (Naive_bayes)	GNB	Non-demented(0)	67	85	75	52	73
			Demented(1)	83	63	72	60	
6.	KNeighbors Classifier	KNN	Non-demented(0)	62	75	68	52	67
			Demented(1)	73	60	66	60	

3.4 Receiver operator characteristic (ROC/AUC) curve

The true positive rate, also known as sensitivity, and false positive rate, also known as specificity, are graphed in the ROC/AUC curve. The sensitivity-specificity pair was indicated by each data point on the ROC/AUC curve. For evaluating the overall classification accuracy of the AD tests, the ROC/AUC curve is the most appropriate graph. The accuracy increases as the curve gets closer to the upper left corner.

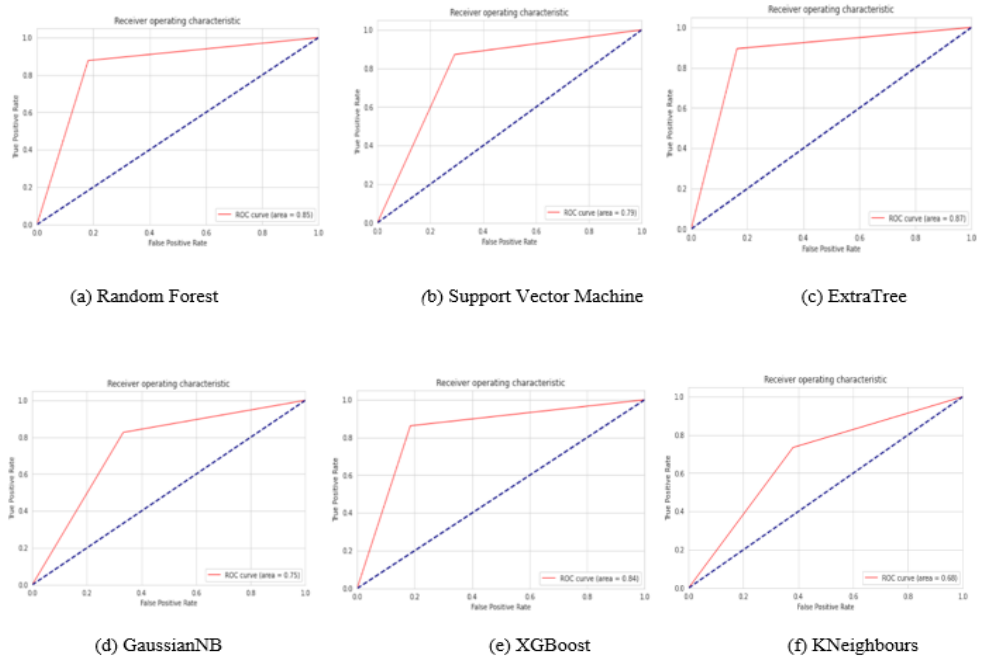


Figure 3. ROC/AUC curve of (a) Random Forest (f) Support Vector Machine (g) ExtraTree (h) GaussianNB (i) XGBoost (j) KNeighbour.

4. CONCLUSION

Since there is currently no known cure for Alzheimer's, it is crucial to lowering risk, offer early intervention, and precisely evaluate symptoms. According to the literature review, multiple attempts have been made to identify Alzheimer's Disease using various machine learning algorithms however, it is still difficult to uncover pertinent characteristics that can identify Alzheimer's extremely early. We presented a system based on supervised learning models for categorising AD patients into two groups, i.e., AD or non-AD based on longitudinal brain MRI data. The six states of art supervised classifiers named Random Forest (RF), XGBoost, Support Vector Machine (SVM), ExtraTree Classifier, Naive Bayes (NB) Classifier, and KNearest Neighbours (KNN) are used to classify and predict. The performance of the ML classifier was evaluated using performance metrics such as precision, recall, accuracy, f1-score and ROC/AUC curve. Among all six ML classifiers, the ExtraTree performed the best with the highest classification accuracy of 87%.

REFERENCES

1. Sandhya Joshi, Vibhudendra Simha G.G., Deepa Shenoy P., Venugopal K.R., Patnaik L.M. “Classification and treatment of different stages of Alzheimer’s disease using various machine learning methods”, International Journal of Bioinformatics Research, ISSN: 0975–3087, Volume 2, Issue 1, 2010

2. Khan, A., & Zubair, S. (2020, July). Expansion of regularized kmeans discretization machine learning approach in the prognosis of dementia progression. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.
3. GBD 2019 Collaborators. Global mortality from dementia: Application of a new method and results from the Global Burden of Disease Study 2019. Alzheimer's Dement (N Y). 2021 Jul 27;7(1): e12200. doi: 10.1002/trc2.12200. PMID: 34337138; PMCID: PMC8315276.
4. Fashtakeh, S. E. (2012). Early Detection of Alzheimer's Disease using Structural MRI: A Research Idea. Life Science Journal, 9(3), 1072-1079. [5] Kavitha C, Mani V, Srividhya SR, Khalaf OI and Tavera Romero CA (2022) Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models. Front. Public Health 10:853294.
5. Battineni, G., Hossain, M. A., Chintalapudi, N., Traini, E., Dhulipalla, V. R., Ramasamy, M., & Amenta, F. (2021). Improved alzheimer's disease detection by MRI using multimodal machine learning algorithms. Diagnostics, 11(11), 2103.
6.] Battineni, G.; Hossain, M.A.; Chintalapudi, N.Traini, E.Dhulipalla, V.R.Ramasamy, M. Amenta, F. Improved Alzheimer's Disease Detection by MRI Using Multimodal Machine Learning Algorithms. Diagnostics 2021, 11, 2103
7. Khan, A., & Zubair, S. (2020, November). A Machine Learning-based robust approach to identify Dementia progression employing Dimensionality Reduction in Cross-Sectional MRI data. In 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH) (pp. 237-242). IEEE.
8. Sisodia, P.S., Ameta, G.K., Kumar, Y. et al. A Review of Deep Transfer Learning Approaches for Class-Wise Prediction of Alzheimer's Disease Using MRI Images. Arch Computat Methods Eng (2023). https://doi.org/10.1007/s11831-022-09870-0
9. Yogesh Kumar, Apeksha Koul, Pushpendra Singh Sisodia, Jana Shafi, Kavita Verma, Mehdi Gheisari, Mohamad Bagher Davoodi, "Heart Failure Detection Using Quantum-Enhanced Machine Learning and Traditional Machine Learning Techniques for Internet of Artificially Intelligent Medical Things", Wireless Communications and Mobile Computing, vol. 2021, Article ID 1616725, 16 pages, 2021. https://doi.org/10.1155/2021/1616725
10. P. S. Sisodia, A. Gupta, Y. Kumar and G. K. Ameta, "Stock Market Analysis and Prediction for Nifty50 using LSTM Deep Learning Approach," *2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)*, Gautam Buddha Nagar, India, 2022, pp. 156-161, doi: 10.1109/ICIPTM54933.2022.9754148.