MACHINE LEARNING GUI BASED FOR DETECTING ALZHEIMER'S

Fatema Nafa[1], Evelyn RodriguezArgueta[1],
Annie Dequit[1] and Changqing Chen[2]

[1]Department of Computer Science, Salem State University, Salem, MA
[2]Department of Chemistry and Physics, Salem State University, Salem, MA

ABSTRACT

Alzheimer's disease (AD), a kind of dementia, is marked by progressive cognitive and behavioural problems that appear in middle or late life. Alzheimer's disease must be detected early in order to create more effective therapies. Dr. Alois Alzheimer was the first doctor in the medical field to notice an unusual state of change in the brains of his deceased patients with mental illness, which marked the start of Alzheimer's study. Machine learning (ML) techniques nowadays employ a variety of probabilistic and optimization strategies to allow computers to learn from vast and complex datasets. Because of the limited number of labelled data and the prevalence of outliers in the current datasets, accurate dementia prediction is extremely difficult. In this research, we propose a sustainable framework for dementia prediction based on ML techniques such as Support Vector Machine, Decision Tree, AdaBoost, Random Forest, and XGmodel. All the experiments, in this literature, were conducted under the same experimental conditions using the longitudinal MRI Dataset.

KEYWORDS

Machine learning, Alzheimer's disease, Feature selection, Biomechanical parameters.

1. INTRODUCTION

The earliest research into the Alzheimer's field began with Dr. Alois Alzheimer who was the first doctor within the medical field to begin to notice an unusual state of change of the brain from his deceased patients with mental illness. This article helped share some light on how the evolution of Alzheimer's research has changed drastically throughout the years. Early onset research focused on the initial connection linked with the genetic traits that were present for the older members of families that had AD and those younger members that over a course of time began showing symptoms or for some generations would never get to that point of progression with the disease [1]. As technology began making its strides forward, we followed its lead and used this new technology to make the discovery that there was a consistent behaviour seen with the high production of amyloid beta and individuals with
Alzheimer's disease.

Human based detection of this disease is a time-consuming and expensive process that requires a large amount of data and the involvement of an experienced clinician.

Automated systems are not prone to human mistake, they are more accurate than human assessments and can be employed in medical decision support systems. Several studies on

David C. Wyld et al. (Eds): AI, AIMLNET, BIOS, BINLP, CSTY, MaVaS, SIGI - 2022
pp. 135-143, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121813

Alzheimer's disease diagnostics have been completed, with the focus recently shifting to the accurate prediction of the disease's early stages.

Furthermore, the current rapid advancement of machine learning technology, which uses AI to anticipate various result, has had a substantial impact on medicine [2], [3]. Recently, several machine learning modelled has been used for AD detection. such as SVM (support vector machines), KNN (K-nearest neighbour), NN (Neural Network) [4]– [6]. However, it remains a major challenge to select best parameters for each model.

In this study, we will build a more effective prediction GUI system for Alzheimer's detection at an earlier stage by using machine learning models, such as Support Vector Machine, Decision Tree, AdaBoost, Random Forest, and XGmodel. The models are compared, CatBoost model clearly outperforms various other models. The results of the experiment reveal that using the CatBoost model for a B-cell epitope prediction with high prediction accuracy, stability, and speed may be developed.

2. RELATED WORK

Many scholars have advocated a comprehensive study on the classification and diagnosis of Alzheimer's disease. A quick review of the relevant work is included in this section.

The major drivers of advanced data analysis model are artificial intelligence (AI) and machine learning (ML). ML approaches are built on the basis of data representations and discover significant insights from the data. The most important step in creating such models is extracting features from the data. Models can be produced manually automatically based on the data representations [2]. The authors [3] utilized a semi-supervised learning model to create a MRI biomarker of MCI-to-AD conversion. Mild cognitive impairment (MCI) is a transitional stage between age-related cognitive decline and Alzheimer's disease (AD). The authors [7] suggested a new eigenbrain-based computer-aided diagnostic (CAD) system for MRI brain imaging. They used kernel support vector machines with multiple kernels to generate the discriminant areas that separate AD from NC, MIE coefficients with values higher than 0.98 were highlighted. Sharma et al. in [8] introduced and tested a support vector machine (SVM) model for discriminating between patients with Alzheimer's disease (AD) and older control subjects of whole-brain anatomical magnetic resonance imaging. Another work by [4] investigated Alzheimer's infections. Using support vector machine (SVM) The researchers studied three main segments: Sagittal, Axial, and Frontal regions of the cerebrum.

The most relevant work is featured in this section, but there are a few others that have investigated the same problem. The scope of this work is limited for biomarkers produced from MRI images. In the above discussed stateof-art techniques are bounded by their performance over many limitations. The authors of this study proposed a machine learning models, Support vector machine, Decision Tree, AdaBoost, Random Forest, and XGmodel to accurately predict progress of a patient from mild cognitive impairment to dementia.

3. PROBLEM DEFINITION

To help clinicians and therapies predicting an early dementia of a patient we propose to develop a machine learning models that can accurately predict progress of a patient from mild cognitive impairment to dementia. Also, the models have a friendly GUI for non-programmers, users do not need advanced knowledge to take full advantage of these models.

4. METHODOLOGY

This section focuses on methodology used for this study. User Graphical Interface (GUI) was developed using Python. GUI designed to predict the dementia. GUI employs easy architecture to be used by non-programmer users. Users can upload their dataset; clean the data using data pre-processing tool; and perform the prediction using multiple machine learning models. The data pre-processing tool provides a set of GUI icons that allow user to read and clean the dataset. Then prediction tool provides icons that can help user to easily use the models to help predict dementia with controlling of different parameters. The section will explain the dataset, data pre-processing, crossfold validation, and Machine Learning models respectively.

4.1. Dataset

The ML models were trained and tested on publicly available longitudinal MRI data[9]. The dataset consists of 373 in total, there are around 39% demented cases in the dataset i.e., majority of the data is of non-Demented cases while 10% of the data is of Converted as shown in Fig.1. The descriptions of the attributes and brief statistical summary are shown in Table. 1.

The paper used the dataset to investigate the following:

a) To assist physicians and therapies in detecting a patient's early dementia b) what biomarkers (or variables) are associated with dementia? c) Is the data of the different datasets normally distributed? d) How would you build and evaluate a predictive model, for dementia on these data?

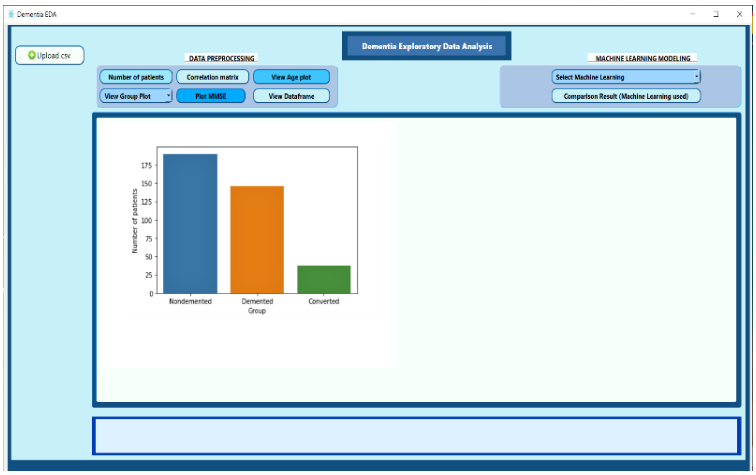

Figure 1. Distrbuation of Dementia Group.

Table 1. The Overview of the Dementia Patient Dataset.

Attribute	Description	Mean±Std
1. Group	Patient grouped as Converted (Previously Normal but developed dimentia later), Demented and Nondemented (Normal Pateints)	1.412±0.6644
2. M/F	Gender	0.571±0.495
3. Age	Age in years	17.134±7.641
4. EDUC	Years of education	5.611±2.59
5. SES	Socioeconomic status as assessed by the Hollingshead Index of Social Position and classified into categories from 1 (highest status) to 5 (lowest status)	1.93±1.55
6. MMSE	Mini-Mental State Examination score (range is from 0 = worst to 30 = best)	15.059±3.84
7. CDR	Clinical Dementia Rating (0 = no dementia, 0.5 = very mild AD, 1 = mild AD, 2 = moderate AD)	0.58±0.72
8. eTIV	Estimated Total Intracranial Volume	142.92± 78.14
9. nWBV	Normalize Whole Brain Volume	65.83± 32.94
10.ASF	Atlas Scaling Factor	132.22 ±71.0056

4.2. Data Pre-Processing

This step enhances data quality by detecting and eliminating mistakes and inconsistencies. GUI in Fig.2. showing the main interface of the system. The left side include the pre-processing part. This consist of checking the overall distribution of categorical and numerical columns, the missing values, outlier rejection, and feature selection of the attribute. The data pre-processing described as follows:

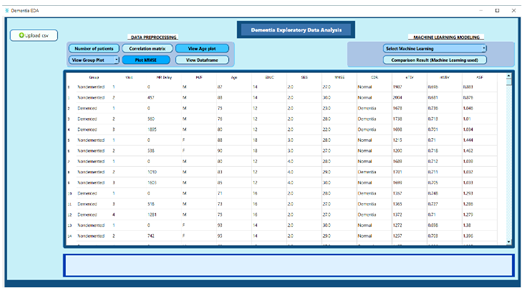

Figure 2:The main GUI of the system.

Outlier detection [10], [11] outliers can negatively affect the training process of the model resulting in lower accuracy. there are different techniques can be used to detect the outliers. In this paper IQR (Inter Quartile Range) Score been used to detect the outliers [12], [13].

Missing values, it is important to identify and manage missing values effectively during data preparation; it led to drawing incorrect conclusions and inferences from the data[14]. There are two approaches to deal with missing data, deleting the rows that has missing values and calculating the mean or median. In this study, calculating the median method is used. Some features have float value cannot impute a float value of mean in place however median imputed, and median is the most representative value of the features in this scenario. Calculating the mean values of the attributes can be formulated as in (1).

$$M(x) = \begin{cases} mean(x) & if\ x = null \\ x & otherwise \end{cases}$$

The standardization the practice of rescaling attributes to achieve a standard normal distribution with zero mean and unit variance is known as standardization or Z-score normalization. As illustrated in Fig.3. standardization (R) reduces the skewness of the data distribution.

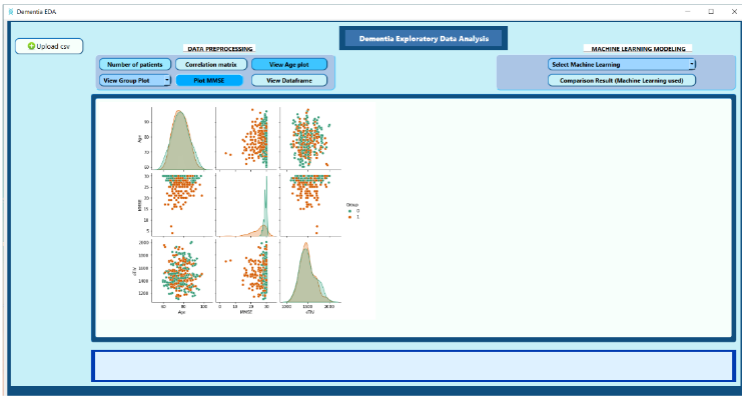

Figure 3. The Population Distribution of all Features.

Feature Engineering, there are approximately 168 Nondemented cases in the dataset, with 34 Converted cases. There are also 97 cases of Dementia. For further information, we investigate another key characteristic called Clinical Dementia Rating. There is a scoring reference used to help doctors determine proper ratings, according to [15], [16]. The score (Normal=0, Very Mild Dementia=0.5, Mild Dementia=1, Moderate Dementia=2, Severe Dementia=3) can be used to characterize and track a patient's level of impairment or dementia.

4.3. Cross-Fold Validation

The K-fold Cross-validation (KCV) technique is one of the most widely used approaches by practitioners for model selection and error estimation of classifiers [15]. The k-fold cross-validation procedure in this study implemented using the scikit-learn machine learning library in Python.

4.4. Machine Learning Models

This is the most crucial step, which includes the development of a dementia prediction model. Various machine learning techniques for dementia prediction have been implemented. The

models are Support Vector machine, Decision Tree, Random Forest, Ada Boost model, and Gradient Boost model. After feeding the input dataset, the model will predict using machine learning techniques and deliver the best result in the form of a comparison to predict the best accuracy for detecting dementia. In this experiment, the best result has been delivered using random forest model. A random forest model is made up of tree-structured [17], [18]. The model adds objects from an array of input to each forest tree. Every single tree vote for classification of the elements of the unit vector independently. The forest selects out the classifications with the greatest votes.

5. RESULT

5.1. Experiment

This section discusses the dataset used in the experiment and the technical requirements. The technical requirements are Python programming language. Then import the necessary libraries and import the dataset to the Jupyter notebook. An experiment was conducted on dataset has been used around 90 blood biomarkers and other demographic data, the total size was 43.8 KB. Table 2. provides a descriptive statistic to describe and summarize the data. It uses quantitative approach describes and summarizes data numerically. We can observe that the data set contain 373 rows and 15 columns. The missing values are 21.

To get a further insight into the data, correlation values were calculated to know how much an attribute affects the dementia attribute (Outcome) or if other attributes are affected by it. Correlation values were calculated using the Pearson (product-moment) correlation coefficient equation. It computes the ratio of the covariance of both features to the product of their standard deviations consequently finding the measure of the linear relationship between those two features. Correlation values are shown in Fig.4.

Table 2. Statistical Information about Dementia Dataset.

Dataset statistics	
Number of variables	15
Number of observations	373
Missing cells	21
Missing cells (%)	0.4%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	43.8 KiB
Average record size in memory	120.3 B

Variable types
Categorical
Numeric

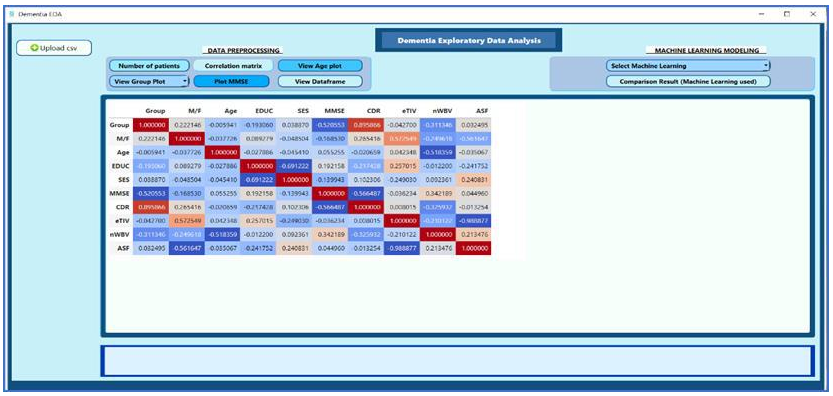

Figure 4. The confusion matrix of the attribute's correlation with the outcome.

Histograms were also created to provide a better visual representation of the data, as seen in Fig.5. Aside from the improved presentation that histograms provide, the figures can make it easier to see potential outliers that could harm the proposed model.

5.2. Evaluation Metrics

The metrics used for measuring the performance of the system are precision, recall, and Accuracy. Here follow the definitions of precision, recall, and Accuracy. see Formulas 1.1,1.2 ,and 1.3 respectively [19], [20].

Precision: Precision is calculated by multiplying the number of TP by the number of TP 'Plus' the number of FP. False positives occur when a model is wrongly classified as positive when it is negative.

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

Recall: the number of true TP separated by the TP '+' FN used to calculate the recall.

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

Accuracy: another metric is accuracy, which is defined as the percentage of true cases (both positive and negative) among all examples retrieved

$$\text{Accuracy} = \text{AUC} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fp} + \text{fn}}$$

5.3. Results Comparison

As it shown in Fig.6. Decision Tree model is very good when we have no idea on the data. Even with unstructured and semi structured data like text, images, and trees Decision Tree algorithm works well. The drawback of the Decision Tree model is that to achieve the best prediction results for any given problem, several key parameters are needed to be set correctly.

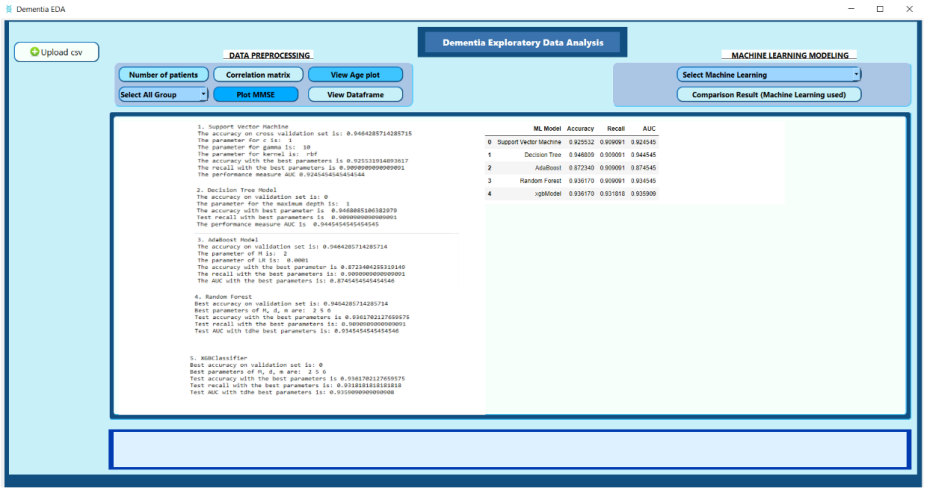

Figure 5. Five Machine learning Models and their Comparison Accuracy of Result.

6. CONCLUSION AND DISCUSSION

One of the real-world medical challenges that must be solved is the early detection of dementia. Throughout this research, deliberate attempts have been made toward developing a framework that will eventually be used to predict illnesses such as dementia. Four machine learning models used; Decision Tree algorithm was explored and assessed on several measures during this project, Decision Tree model achieved the highest accuracy. The goal of this project is to create a system that can predict dementia in a patient earlier and more accurately utilizing machine learning techniques that provide advance support for dementia prediction accuracy.

REFERENCES

[1] L. Caly, J. D. Druce, M. G. Catton, D. A. Jans, and K. M. Wagstaff, "The FDA-approved drug ivermectin inhibits the replication of SARS-CoV-2 in vitro," Antiviral Res., vol. 178, p. 104787, 2020.

[2] S. Sharma and P. K. Mandal, "A Comprehensive Report on Machine Learning-based Early Detection of Alzheimer's Disease using Multi-modal Neuroimaging Data," ACM Comput. Surv. CSUR, vol. 55, no. 2, pp. 1–44, 2022.

[3] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, J. Tohka, and A. D. N. Initiative, "Machine learning framework for early MRIbased Alzheimer's conversion prediction in MCI subjects," Neuroimage, vol. 104, pp. 398–412, 2015.

[4] A. Sharma, S. Kaur, N. Memon, A. J. Fathima, S. Ray, and M. W. Bhatt, "Alzheimer's patients detection using support vector machine (SVM) with quantitative analysis," Neurosci. Inform., vol. 1, no. 3, p. 100012, 2021.

[5] T. A. Assegie, "Support Vector Machine And K-Nearest Neighbor Based Liver Disease Classification Model," Indones. J. Electron. Electromed. Eng. Med. Inform., vol. 3, no. 1, pp. 9–14, 2021.

[6] M. A. Ibrahim, M. U. G. Khan, F. Mehmood, M. N. Asim, and W. Mahmood, "GHS-NET a generic hybridized shallow neural network for multi-label biomedical text classification," J. Biomed. Inform., vol. 116, p. 103699, 2021.

[7] Y. Zhang et al., "Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning," Front. Comput. Neurosci., vol. 9, p. 66, 2015.

[8] B. Magnin et al., "Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI," Neuroradiology, vol. 51, no. 2, pp. 73–83, 2009.

[9] J. H. Kramer et al., "Longitudinal MRI and cognitive change in healthy elderly.," Neuropsychology, vol. 21, no. 4, pp. 412– 418, 2007, doi: 10.1037/0894-4105.21.4.412.

[10] I. Ben-Gal, “Outlier detection,” in Data mining and knowledge discovery handbook, Springer, 2005, pp. 131–146.
[11] V. Hodge and J. Austin, “A survey of outlier detection methodologies,” Artif. Intell. Rev., vol. 22, no. 2, pp. 85–126, 2004.
[12] S. Walfish, “A review of statistical outlier methods,” Pharm. Technol., vol. 30, no. 11, p. 82, 2006.
[13] C. Andreou and V. Karathanassi, “Estimation of the number of endmembers using robust outlier detection method,” IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 7, no. 1, pp. 247–256, 2013.
[14] W. Z. Liu, A. P. White, S. G. Thompson, and M. A. Bramer, “Techniques for dealing with missing values in classification,” in International Symposium on Intelligent Data Analysis, 1997, pp. 527–536.
[15] J. C. Morris, “Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the Alzheimer type,” Int. Psychogeriatr., vol. 9, no. S1, pp. 173–176, 1997.
[16] W. J. Burke et al., “Reliability of the Washington University clinical dementia rating,” Arch. Neurol., vol. 45, no. 1, pp. 31– 32, 1988.
[17] C. Brokamp, R. Jandarov, M. Hossain, and P. Ryan, “Predicting daily urban fine particulate matter concentrations using a random forest model,” Environ. Sci. Technol., vol. 52, no. 7, pp. 4173–4179, 2018.
[18] S. J. Rigatti, “Random forest,” J. Insur. Med., vol. 47, no. 1, pp. 31–39, 2017.
[19] H. Dalianis, “Evaluation metrics and evaluation,” in Clinical text mining, Springer, 2018, pp. 45–53.
[20] J.-O. Palacio-Niño and F. Berzal, “Evaluation metrics for unsupervised learning algorithms,” ArXiv Prepr. ArXiv190505667, 2019.