Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/74/20240455

Statistical analysis and prediction of data for Alzheimer's disease and mild cognitive impairment based on multiple machine learning algorithms

Zijun Lan

College of science and technology hnut, Hunan University of Technology, Hunan, Zhuzhou 412008, China

Corresponding author email: 1278090081@qq.com

Abstract. This paper focuses on the characteristics and diagnostic methods of two neurodegenerative diseases, Alzheimer's disease and mild cognitive impairment, as well as the application of machine learning in the medical field. Firstly, through the statistical analysis of patients' age and gender, it is found that patients with Alzheimer's disease tend to be older than patients with mild cognitive impairment, while the number of female patients are both larger than that of male. These results provide an important basis for subsequent classification prediction. Next, this paper uses a variety of machine learning algorithm models for prediction, including Decision Tree, Random Forest, Logistic Regression, Naive Bayes, KNN, and SVM. These models are all based on different algorithmic principles and data processing methods to achieve classification prediction of patients with Alzheimer's disease and mild cognitive impairment. By comparing the prediction accuracy of each model, it is found that Decision Tree and Naive Bayes models have the best prediction effect, with a prediction accuracy of 75%; while Random Forest and Logistic Regression models also show high prediction accuracy. In contrast, the KNN and SVM models had a poor prediction effect of 69.2%. These results suggest that the use of decision tree or Naive Bayes models for prediction should be prioritised when classifying patients with Alzheimer's disease and mild cognitive impairment. Finally, this paper highlights the importance of machine learning in the medical field. By analysing and processing large amounts of medical data, machine learning can help doctors diagnose diseases more accurately and provide more effective treatment options. The application of these techniques will help improve the quality of life of patients and advance the field of medicine.

Keywords: Alzheimer's disease, Mild cognitive impairment Machine learning algorithms.

1. Introduction

Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI) are neurodegenerative disorders, and the definitive diagnosis of these diseases requires a variety of examinations and tests [1]. In recent years, machine learning techniques have been increasingly used in the medical field [2]. Machine learning can analyse large amounts of data and assist doctors in diagnosis and treatment through pattern recognition and prediction. This paper will explore the role of machine learning in diagnosing Alzheimer's disease and mild cognitive impairment [3].

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/74/20240455

Firstly, machine learning can help doctors to analyse a large amount of clinical data, including information such as patient's age, MMSE score, CSF biomarkers and APOE4 alleles. This data can be used to train machine learning algorithms and build predictive models. With these models, doctors can more accurately predict whether a patient has AD or MCI [4,5].

Second, machine learning can also help doctors discover new biomarkers and determine their relationship with AD or MCI. One researcher used machine learning techniques to analyse a large amount of brain imaging data and discovered a new type of biomarker called "brain morphological features" [6]. These features can be used to predict whether a patient has AD or MCI and help diagnose these disorders earlier [7].

In addition, machine learning can help doctors determine the most effective treatments [8]. One researcher used machine learning techniques to analyse a large amount of clinical data and discovered a new treatment that slows the progression of AD by modulating the immune system [9]. This study shows that machine learning can help doctors discover new treatments and improve the effectiveness of treatment [10].

In conclusion, machine learning has an important role in diagnosing Alzheimer's disease and mild cognitive impairment. By analysing large amounts of clinical data and biomarker data, machine learning can build accurate predictive models and help doctors diagnose these neurodegenerative diseases earlier. In addition, machine learning can help doctors discover new biomarkers and treatments, and improve treatment outcomes. In this paper, we present statistical analysis and prediction of data on Alzheimer's disease and mild cognitive impairment based on multiple machine learning methods to help doctors better diagnose and treat patients with AD and MCI.

2. Introduction to the dataset

The dataset used in this paper is a publicly available dataset from UCL and includes a number of indicators such as patient's age, MMSE scores, cerebrospinal fluid biomarkers and APOE4 alleles, with the categorical variable being the patient's type of disease (Alzheimer's disease versus mild cognitive impairment).

Age is an essential factor because neurodegenerative diseases usually show an age-related pattern. By analysing the age differences between the ADD and MCI groups, it was possible to gain insight into the age dependence of these diseases.

The MMSE is a widely used cognitive screening tool. By examining the MMSE scores of patients with AD and MCI, it is possible to get a comprehensive picture of cognitive decline and to help differentiate the severity of impairment between these two diagnostic groups.

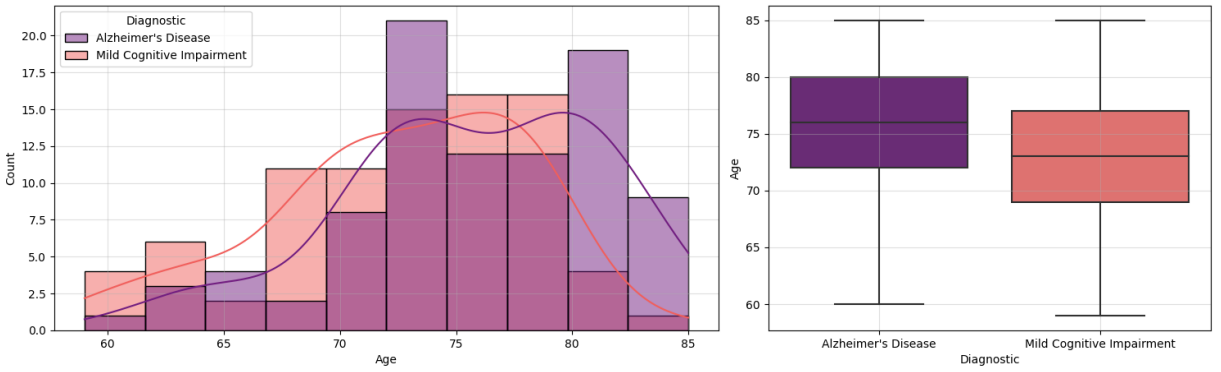
Cerebrospinal fluid biomarkers are critical to understanding the neurochemical changes associated with AD and MCI. Amyloid concentrations reflect plaque accumulation, whereas total and phosphorylated Tau proteins indicate neurofibrillary tangle formation. The study of these biomarkers has helped to identify the unique biochemical features of AD and MCI.

The APOE4 allele has been associated with an increased risk of AD. By analysing its presence in AD and MCI, its potential role as a genetic marker in differentiating between these two diagnostic categories can be explored.

3. Statistical analysis of data

The histogram statistics of age of patients with Alzheimer's disease and mild cognitive impairment were plotted as shown in Fig. 1.(a), box plots of patients with Alzheimer's disease and mild cognitive impairment were plotted as shown in Fig. 1.(b), and the statistical results of age are shown in Table 1.

DOI: 10.54254/2755-2721/74/20240455



(a) (b)

Figure 1. (a) Age histogram, (b) Age box-line chart.
(Photo credit: Original)

Table 1. Model evaluation.

Diagnostic	count	mean	std	min	25%	50%	75%	max
Alzheimer's Disease	91	75.6	5.6	60	72	76	80	85
Mild Cognitive Impairment	86	72.5	5.6	59	69	73	77	85

From the results, it can be seen that the age of patients with Alzheimer's disease is mainly concentrated at 75 years old, and the age of patients with mild cognitive impairment is mainly concentrated at 72 years old, and the distribution shows that patients with Alzheimer's disease tend to be older than patients with mild cognitive impairment.

The statistics of patients with Alzheimer's disease and mild cognitive impairment in different age groups according to gender are shown in Figure 2.

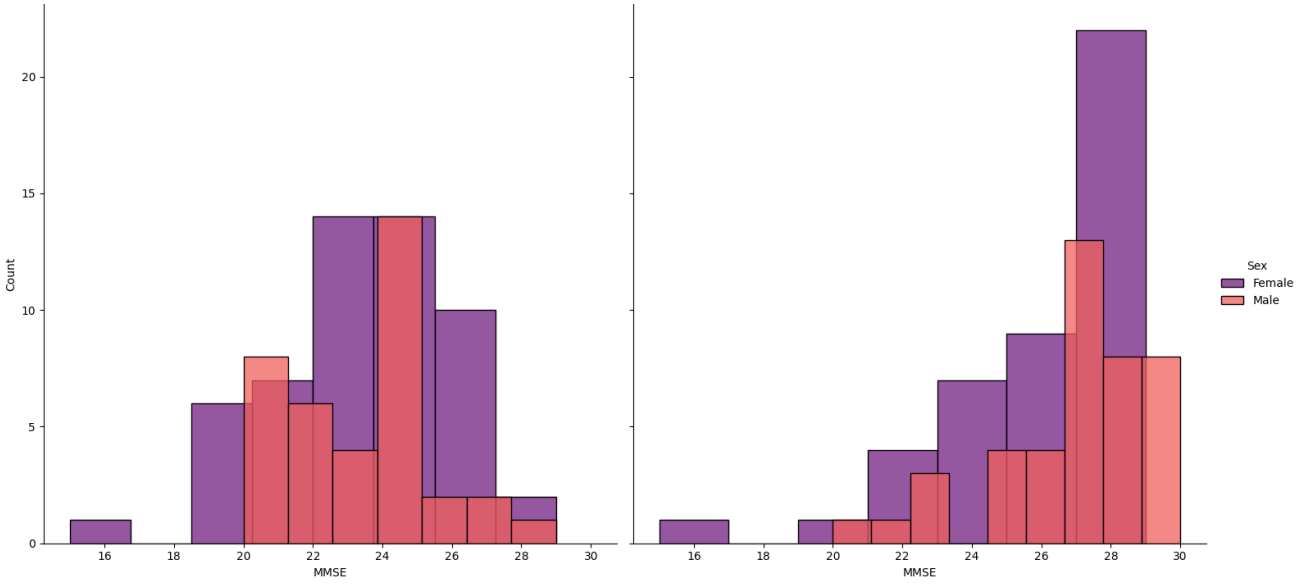


Figure 2. Histogram of gender statistics.
(Photo credit: Original)

DOI: 10.54254/2755-2721/74/20240455

As can be seen in Figure 2, the number of female patients with both Alzheimer's disease and mild cognitive impairment is greater than that of males in all age groups.

4. Machine Learning Algorithms

4.1. Decision tree

Decision tree (dt) is a classification model based on a tree structure, which constructs a tree structure by partitioning the data, where each node represents a feature attribute, each branch represents a different value of that attribute, and the leaf nodes represent a category. The commonly used algorithms for decision trees are ID3, C4.5 and CART.

4.2. Random Forest

Random Forest (RF) is an integrated learning method that consists of multiple decision trees, each independently sampled from the data. When classifying, Random Forest votes on the results of each decision tree and finally selects the category with the most votes as the classification result.

Logistic regression

Logistic regression (LR) is a linear model widely used in binary classification problems, which predicts the output probability by multiplying the input features with weights and adding a bias term mapped to between 0 and 1 by a sigmoid function. Logistic regression can be trained using optimisation algorithms such as gradient descent.

4.3. Naive Bayes

Naive Bayes (NB) is a classification method based on Bayes' theorem and the assumption of feature independence. It assumes that each feature is independent from other features and calculates the posterior probability of each category based on the probability of the feature's occurrence, and finally selects the category with the largest posterior probability as the classification result. Commonly used algorithms for Simple Bayes are Gaussian Simple Bayes, Polynomial Simple Bayes and Bernoulli Simple Bayes.

4.4. K Nearest Neighbour

K Nearest Neighbour (KNN) is an instance-based learning method that determines the category to which a new sample belongs by calculating the distance between the new sample and all the samples in the training set. Specifically, KNN selects the K samples in the training set that have the closest distance to the new sample and votes according to the categories to which these K samples belong, and finally selects the category with the most votes as the classification result.

4.5. Support Vector Machine

Support Vector Machine (SVM) is a binary classification model that achieves classification by mapping data into a high-dimensional space and finding an optimal hyperplane in that space.SVMs can use different types of kernel functions to achieve nonlinear classification and optimise the model's performance by adjusting hyperparameters such as penalty coefficients and kernel function parameters.

5. Experiments and Results

After the statistical analysis of the data of patients with Alzheimer's disease and mild cognitive impairment, this paper classifies patients with Alzheimer's disease and mild cognitive impairment based on machine learning algorithms models such as Decision Tree, Random Forest, Logistic Regression, Naive Bayes, KNN and SVM. The training set, validation set and test set were divided according to the ratio of 6:2:2, and the prediction effect of the model was evaluated using Accuracy, Precision, Recall and F1 Score, and the results are shown in Figure 3 and Table 2.

DOI: 10.54254/2755-2721/74/20240455

Table 2. Modelling evaluation.

	DT	RF	LR	NB	KNN	SVM
Accuracy	75.0%	72.2%	72.2%	75.0%	69.4%	69.4%
Precision	75.1%	72.5%	72.5%	79.5%	71.1%	70.0%
Recall	75.0%	72.2%	72.2%	75.0%	69.4%	69.4%
F1 Score	75.0%	72.1%	72.1%	74.0%	68.8%	69.2%

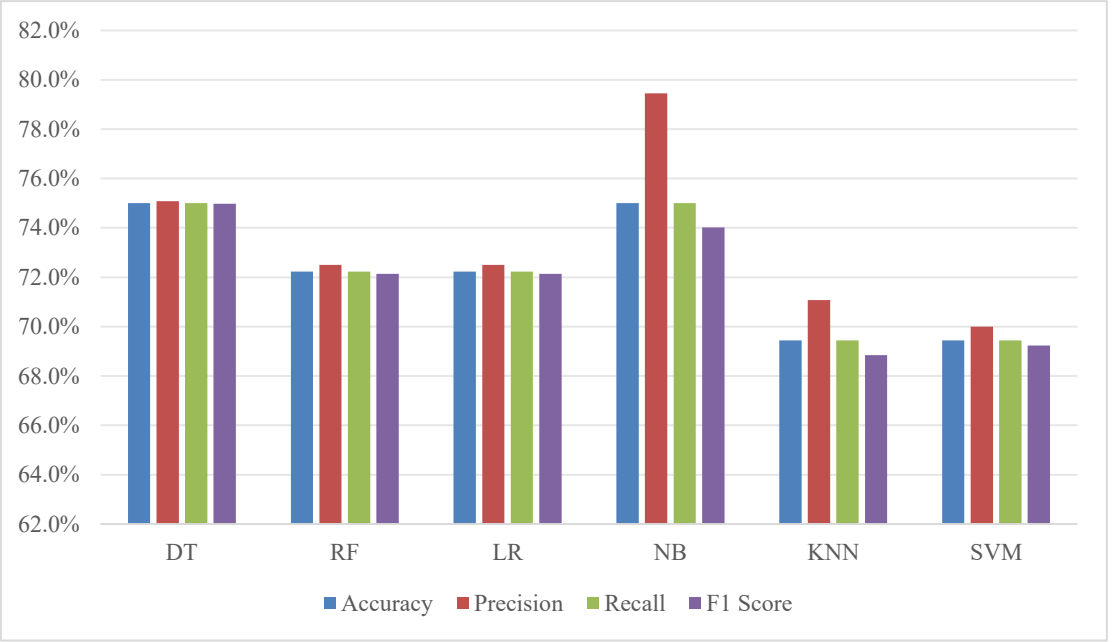


Figure 3. Modelling evaluation.
(Photo credit: Original)

As can be seen from the results, the best prediction results in the classification of patients with Alzheimer's disease and mild cognitive impairment were obtained from the Decision Tree Model and the Naive Bayes Model, both of which achieved 75% prediction accuracy; followed by the Random Forest Model and the Logistic Regression Model, both of which also achieved more than 70% prediction accuracy; and the KNN and SVM Models had a worse prediction effect of only 69.2%.

6. Conclusion

By statistically analysing the age and gender of patients with Alzheimer's disease and mild cognitive impairment, this paper found that patients with Alzheimer's disease tended to be older than those with mild cognitive impairment, while the number of female patients was greater than that of males in all cases. In addition, machine learning algorithm models such as Decision Tree, Random Forest, Logistic Regression, Naive Bayes, KNN and SVM were used to classify patients with Alzheimer's Disease and Mild Cognitive Impairment in this paper and the prediction accuracies of each model were derived. Among them, the best prediction results in the classification of patients with Alzheimer's disease and mild cognitive impairment are the Decision Tree model and Naive Bayes model, both of which have a prediction accuracy of 75%; followed by the Random Forest model and the Logistic Regression model, which both have prediction accuracies of more than 70% as well; and the KNN and SVM models have a poorer prediction effect of only 69.2%.

Taken together, this paper provides a certain reference for the diagnosis of Alzheimer's disease and mild cognitive impairment in the medical field through the statistical analysis of data and machine

DOI: 10.54254/2755-2721/74/20240455

learning algorithm model building. At the same time, this paper also shows the application prospects of machine learning technology in the medical field, providing new ideas and methods for further research and exploration of the treatment and prevention of neurodegenerative diseases.

References

[1] Yi Z ,Shasha S ,Xiaokai L , et al.A multiclass extreme gradient boosting model for evaluat ion of transcriptomic biomarkers in Alzheimer's disease prediction.[J].Neuroscience letter s,2023,821137609-137609.

[2] Qiankun Z, Yanyan S, Ning Z, et al.Alzheimer's Disease Prediction via Brain Structural-Functional Deep Fusing Network.[J].IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society,2023,PP.

[3] Haoliang D ,Huabin W ,Yonglin C , et al.EAMNet: an Alzheimer's disease prediction model based on representation learning.[J].Physics in medicine and biology,2023,68(21):

[4] Preethi T ,Yuvaraj N ,K.R. P S .EAD-DNN: Early Alzheimer's disease prediction using deep neural networks[J].Biomedical Signal Processing and Control,2023,86(PB):

[5] Yiran X ,Longfei M ,Hui Z , et al.Double-attention Assisted Multi-task Learning for the Alzheimer's Disease Prediction from Mild Cognitive Impairmentsup/.[J].Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference,2023,20231-4.

[6] Zahra S ,Omar H K ,Morteza S .Data-driven personalized assortment optimization by considering customers' value and their risk of churning: Case of online grocery shopping[J]. Computers Industrial Engineering,2023,182.

[7] Raymond X G ,Marion C ,Ke Q , et al.Explainable machine learning aggregates polygenic risk scores and electronic health records for Alzheimer's disease prediction.[J].Scientific reports,2023,13(1):450-450.

[8] Zeeshan A ,Hilal T ,To K C .Alzheimer's disease prediction based on continuous feature representation using multi-omics data integration[J].Chemometrics and Intelligent Laboratory Systems,2022,(prepublish):104536-.

[9] Hao G ,Chaoyue W ,Dacheng T .MRI-based Alzheimer's disease prediction via distilling the knowledge in multi-modal data.[J].NeuroImage,2021,244118586-118586.

[10] M. C S ,Jianfeng W ,Jie Z , et al.Improved Prediction of Imminent Progression to Clinically Significant Memory Decline Using Surface Multivariate Morphometry Statistics and Sparse Coding[J].Journal of Alzheimer's Disease,2021,81(1):209-220.