Building a Diagnostic Standard for Alzheimer's Disease Based on Decision Trees and Analyzing Key Factors with Random Forests

Yang Tang

School of Data Science, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China

122090504@link.cuhk.edu.cn

Abstract. Research Background and Significance: Alzheimer's disease (AD) is a widespread neurodegenerative disorder that poses a significant threat to the health of millions of older adults worldwide. Despite the availability of various methods to assess the severity of dementia, the need for a high-precision diagnostic model remains crucial to enhancing patient outcomes. Accurate diagnosis is essential not only for the well-being of individuals but also for the effective management and treatment of the disease. This study aims to address this critical need by developing a more precise diagnostic framework for AD, utilizing advanced machine learning techniques in combination with comprehensive clinical data. Study Contributions: In this research, a decision tree model was constructed based on the principle of information gain, using a meticulously pre-processed sample of 2,149 patients from a public dataset. The model achieved a diagnostic accuracy of 93.47%, markedly outperforming traditional diagnostic methods. Additionally, a random forest model was employed to identify key risk factors influencing AD, such as age and lifestyle habits. These findings not only equip clinicians with more accurate diagnostic tools but also provide a robust scientific foundation for developing AD prevention and treatment strategies. The study also acknowledges its limitations and suggests directions for future research to further improve the diagnosis and understanding of Alzheimer's disease.

Keywords: Machine learning, Decision tree, Random forest, Classifying.

1. Introduction

Alzheimer's disease (AD) is a severe neurodegenerative disorder that has become an increasingly critical focus in global health. Currently, more than 6.7 million older adults in the United States have been diagnosed with AD dementia, and this number is projected to nearly double to 13.8 million by 2060 [1]. In 2019 alone, 121,499 people in the U.S. succumbed to AD, making it the sixth leading cause of death in the country. Alarmingly, the prevalence of AD continues to rise, with the number of deaths attributed to the disease increasing by over 145% between 2000 and 2019 [2]. AD typically leads to neuronal death or dysfunction, resulting in impaired memory, cognitive abnormalities, and even loss of self-care abilities. This progression not only robs patients of their precious memories but also imposes a significant burden on family caregivers, the healthcare system, and society at large. In 2022, more than 11 million individuals provided approximately 18 billion hours of care to those with Alzheimer's disease [3].

Currently, various indicators are used in the medical field to assess the degree of dementia, including the Mini-Mental State Examination (MMSE), Montreal Cognitive Assessment (MoCA), and Hamilton Depression Rating Scale (HAMD). However, these individual indicators cannot serve as definitive diagnostic criteria for Alzheimer's disease, highlighting the urgent need for more accurate diagnostic models [4]. Machine learning is increasingly being employed in medical applications, and among the various models and algorithms, decision trees are highly valued for their interpretability and ease of use, making them particularly suitable for medical decision-making [5]. The aim of this study is to leverage decision tree algorithms to innovatively integrate multiple evaluation indicators and improve the accuracy of AD diagnosis.

In addition, this study introduces a novel two-layer cyclic algorithm designed to identify the key factors influencing AD, such as age, environment, and lifestyle choices [6]. The findings of this

research are not only applicable to the direct diagnosis of AD but also provide valuable insights for future research directions in this field [7].

Contributions of this study include:

- Development of an innovative diagnostic model for AD that integrates multiple evaluation indicators using decision tree algorithms.
- Introduction of a two-layer cyclic algorithm to identify and analyze the key factors affecting AD, such as age, environmental factors, and lifestyle choices.
- Providing a more accurate diagnostic tool for clinicians, potentially improving patient outcomes.
- Offering a scientific foundation for future research in AD diagnosis and treatment.
- Identifying potential areas for further exploration, as discussed in the paper's conclusion and future research directions.

2. Relevant Theories

2.1. Definition and Background of Alzheimer's Disease

Alzheimer's disease, mainly occurs in the elderly. It is a kind of primary degenerative encephalopathy of the central nervous system, which is a kind of persistent high-level neurological activity dysfunction, that is, memory, thinking, analysis and judgment, visual spatial recogni, ion, emotion and other disorders in the absence of consciousness disorder. The onset of AD is insidious and complex etiology, and its occurrence is the result of the interaction of many factors, including genetic, lifestyle and environmental factors [8]. There are no specific therapeutic drugs and means to treat or reverse the progression of the disease. Nonetheless, studies have found that Alzheimer's disease causes amyloid deposition in the brain and fibrillary tangles in neurons [9]. Ultimately, because neural cells "silently" atrophy or even die, or have abnormal intercellular signaling, they lead to cognitive dysfunction, such as memory, language, computation, and behavior. Eventually life-threatening.

2.2. Decision Tree Model

Decision tree is a commonly used machine learning model. First, the initial data needs to be accurately classified through a series of tests with different nodes and branches, while the root node covers all the examples [10]. When the test sample passes through a node, the decision tree will be transferred to different branches based on the specific properties of the sample. After passing through multiple nodes and branches, the sample will eventually flow to a leaf node. Therefore, the image is named as the decision tree.

The decision tree divides the sample of each leaf node into a unified class. When building decision trees, we typically classify samples based on the "purity" of the nodes (i.e., the homogeneity of the samples in the nodes). Whereas, purity is usually measured using entropy. Entropy is a measure of randomness in a dataset and is defined as:

$$\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

Where m means the number of all the classes, $\boldsymbol{p_i}$ stands for the proportion of the class I in the dataset D. The goal of constructing a decision tree is to reduce the entropy with each split. The entropy of the data classification segmentation is 0, which is perfect. Decision tree is to ensure the effectiveness of each step of classification by comparing the amount of information gain in different segmentation methods. By calculating the entropy difference before and after the data set segmentation, the data segmentation method with the highest information gain is obtained, which is used as the data segmentation criterion for this node. The formula for the information gain IG for dataset D and attribute A is as follows:

$$\mathrm{IG}(D, A) = \mathrm{Info}(D) - \sum_{j=1}^{v} \frac{|D_j|}{|D|} \mathrm{Info}(D_j) \quad (2)$$

Where v is the number of distinct values in attribute A, $\boldsymbol{D_j}$ is the subset of D for which attribute A has value j, and $|\boldsymbol{D_j}|$ and |D| are the cardinalities of $\boldsymbol{D_j}$ and D, respectively.

2.3. Random Forest Model

Random forest is an ensemble learning model. Compared to a single decision tree model, a random forest model will build hundreds or thousands of decision trees. Each of these trees is trained using a randomly selected subset of data and a subset of features. Eventually, the random forest will integrate the predictions for each tree and come up with a relatively compromise result.

The key advantage of the random forest model is that it solves the overfitting problems which may occur frequently with a single decision tree model. In addition, it is able to assess the importance of individual features to the model's predictions. This is undoubtedly a good way to deal with the problem of feature selection.

3. Experiments and Analysis

3.1. Dataset and Preprocessing

The study data are obtained from the Kaggle website (publicly accessible). The dataset consists of 2,149 rows and 35 columns. Each row represents one person, and each column represents a variable. Contains demographic information: gender, age, race, and education level. Lifestyle factors are also included, such as BMI, smoking, alcohol consumption, physical activity, diet and sleep quality. Comorbidities are also counted, including the family history of AD, cardiovascular disease, diabetes, depression, head injury, and hypertension. Clinical indicators, like systolic and diastolic blood pressure, and cholesterol levels are also included in the dataset. Cognitive and functional measures like MMSE, Functional Assessment, memory complaints, behavioral problems, and ADL scores are also included. The dataset also recorded some psychological symptoms, like confusion, disorientation, personality changes, having difficulty completing tasks, and forgetfulness. Meanwhile, the dataset also contains diagnostic information for Alzheimer's disease (0 being no and 1 being yes) and placeholders for confidential information about the attending physician.

3.2. Diagnostic with Decision Tree Model

3.2.1 Model building

To build a reliable diagnostic decision tree model, the dataset was first divided into two subsets: training data and test data, with a ratio of 8:2. Of these, 80% of the data was used for model training, while the remaining 20% was utilized to evaluate the classification performance. During the training process, five variables were selected for cognitive assessment: functional assessment, MMSE, ADL, behavioral problems, and memory complaints. The decision tree model was designed to be scalable, allowing for further segmentation points until there was no significant improvement in performance on the validation set. The final decision tree model, as shown in Fig 1, can be directly utilized as a basis for clinical diagnosis.

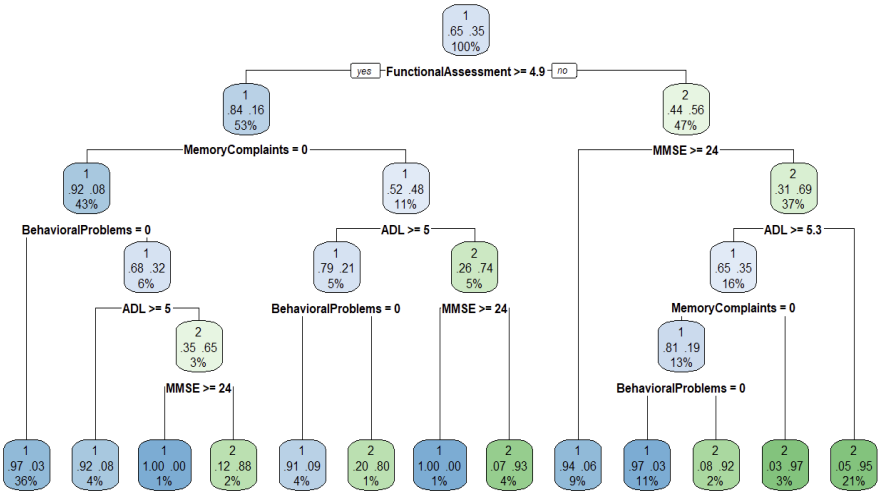

Figure 1. Structure diagram (Photo credit: Original)

In this schematic, all patients are initially categorized based on their Functional Assessment score. If the score exceeds 4.9, the focus shifts to the MMSE score; otherwise, attention is directed to memory issues. For instance, consider a patient with a Functional Assessment index of 5.6, an MMSE score of 32, and an ADL index of 3.3, with no memory or behavioral problems. This patient would first be categorized into the group corresponding to the higher Functional Assessment index. Subsequently, the patient would be further classified into the appropriate node due to the MMSE score being above 24. At the next node, with an ADL score of less than 5.3, the patient's memory and behavioral performance are then evaluated. Ultimately, the decision tree classifies this patient as a healthy individual.

3.2.2 Results analysis

The evaluation of the decision tree model showed a specificity of 93.7% and a sensitivity of 93.1%, respectively. The overall accuracy was as high as 93.5%, indicating a strong consistency between the prediction results and the actual diagnosis. Meanwhile, the ROC curve also further validated the diagnostic performance of this model. An AUC value of 0.94 and high AUC values indicate high diagnostic power and the high ability of the model to distinguish between positive AD and negative cases. As show in the fig 2.

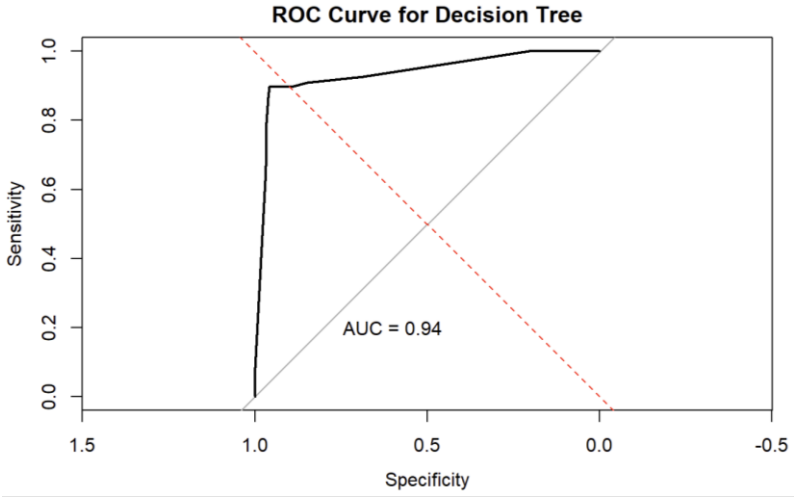

Figure 2. Experimental results (Photo credit: Original)

3.3. Key Factor Analysis Based on Random Forest Model

3.3.1 Model building

Given the small correlation between relevant variables and final results in the dataset, a brand new algorithm was introduced in this study to develop the optimal AD prediction model. This algorithm

employs a two-layer nested loop, where all variables are first disordered into a queue. Subsequently, variables are added one by one to the random forest model based on the training set, with evaluation performed on the test set using the AUC index. Variables that contribute to an increase in AUC are retained, while those that do not are excluded as noise, effectively preventing overfitting. This process is illustrated in Fig 3.

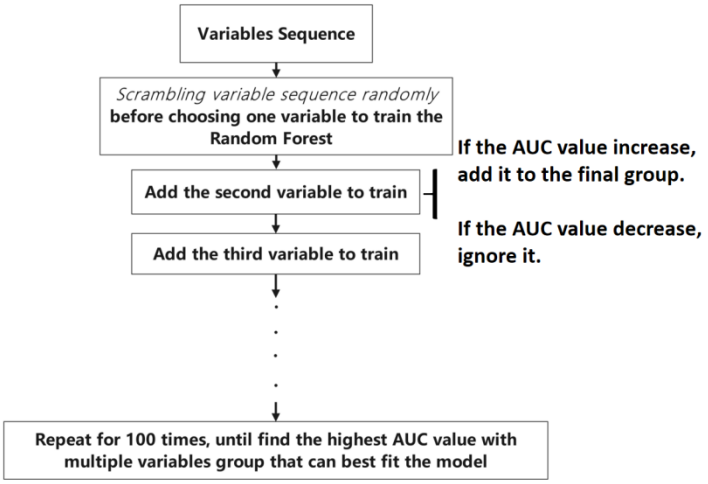

Figure 3. Flow chart (Photo credit: Original)

3.3.2 Results analysis

The algorithm was implemented after 100 shuffling iterations of the variable sequence, successfully identifying a set of variables that significantly enhance the predictive power of the model. The selected variables include age, systolic blood pressure (SystolicBP), cardiovascular disease, drinking habits, family history of Alzheimer's disease, and physical activity. A random forest model incorporating these variables achieved an AUC of 0.639, demonstrating a modest yet meaningful improvement in predictive accuracy compared to random guessing.

4. Challenges and Future Research Directions

In the research, there exist some challenges and unsolved problems. First is about the data source for this study. This dataset has a limited scale. Meanwhile, the data is of low quality due to its poor correlation for all variables and disease outcomes. As for the decision tree, our model achieved a 93.5% accuracy rate. However, assessing Alzheimer's disease is not only about the accuracy of the model, but also about the sensitivity of the model. In practice, we would rather have some misdiagnoses than miss any real case. This is because the missing diagnosis of Alzheimer's disease leads to the loss of timely treatment and intervention, which seriously affects their reasonable treatment.

Therefore, in future research, researchers should pay more attention to data collection, and expand the size and diversity of data sets to cover more regions and classes, so as to improve the diversity and universality of data sources. Researchers should also focus on actual clinical trials to ensure that the data is accurate. When it comes to models, future researchers should pay more attention to the sensitivity of the model and make a decision tree model of more practical value.

5. Conclusion

The project successfully established a decision process system using five assessment indicators, resulting in a decision tree system with a 93.47% accuracy rate. This system enables doctors to determine with relative ease whether a potential Alzheimer's patient actually has the disease, offering a simpler alternative to the more complex random forests and regression models, which are challenging to calculate and use for decision-making purposes.

Furthermore, a double-layer looping algorithm was designed to identify the optimal random forest model for predicting Alzheimer's susceptibility. This model, with an AUC value of 0.639, predicts the likelihood of developing Alzheimer's based on factors such as age, drinking habits, cardiovascular diseases, family history of Alzheimer's, and exercise habits. While the accuracy of this model is not yet sufficient to serve as a medical diagnostic criterion, it demonstrates significant potential for disease prevention. By identifying susceptible individuals, this model allows for targeted tracking and study of their lifestyle habits and health conditions, contributing to a better understanding and prediction of Alzheimer's disease. Additionally, it can assist healthcare institutions and researchers in more effectively allocating resources for early intervention and prevention in high-risk groups.

The model shows promise for further refinement and expansion, particularly in improving its accuracy and applicability as a diagnostic tool. Future research could explore the integration of additional variables and advanced machine learning techniques to enhance predictive capabilities. Moreover, longitudinal studies focusing on the identified high-risk individuals could provide deeper insights into the progression and prevention of Alzheimer's disease. The ultimate goal is to develop a comprehensive and reliable model that can be widely adopted in clinical settings for early detection and prevention, thus contributing to better outcomes for individuals at risk of Alzheimer's disease.

References

[1] Edward Jenner Tettevi, Deryl Nii Okantey Kuevi, Balagra Kasim Sumabe, et al. In Silico Identification of a Potential TNF-Alpha Binder Using a Structural Similarity: A Potential Drug Repurposing Approach to the Management of Alzheimer's Disease. BioMed Research International, 2024, 1: 2024.

[2] Knopman DS, Amieva H, Petersen RC, et al. Alzheimer disease. Nat Rev Dis Primers, 2021, 7: 33.

[3] Ding Yiming, et al. A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using 18F-FDG PET of the Brain. Radiology, 2018, 18 (4): 1598-1695.

[4] Zhu, X., Huang, Y., Wang, X., Wang, R.: 'Emotion recognition based on brain-like multimodal hierarchical perception', Multimedia Tools and Applications, 2024, 83, (18), pp. 56039-56057.

[5] Tanveer M, Richhariya B, Khan RU, et al. Machine learning techniques for the diagnosis of Alzheimer's disease: A review. ACM Trans. Multimedia Comput. Commun. Appl., 2020, 16 (1s): 28.

[6] Zhu X, Guo C, Feng H, et al. A Review of Key Technologies for Emotion Analysis Using Multimodal Information. Cognitive Computation, 2024: 1-27.

[7] Al-Dlaeen D, Alashqur A. Using decision tree classification to assist in the prediction of Alzheimer's disease. In 2014 6th International Conference on Computer Science and Information Technology (CSIT), 2014: 122-126.

[8] Song M, Jung H, Lee S, et al. Diagnostic classification and biomarker identification of Alzheimer's disease with random forest algorithm. Brain Sciences, 2021, 11 (4): 453.

[9] Sun D, Peng H, Wu Z. Establishment and analysis of a combined diagnostic model of Alzheimer's disease with random forest and artificial neural network. Frontiers in Aging Neuroscience, 2022, 14: 921906.

[10] Dimitriadis SI, Liparas D, Alzheimer's Disease Neuroimaging Initiative. How random is the random forest? Random forest algorithm on the service of structural imaging biomarkers for Alzheimer's disease: from Alzheimer's disease neuroimaging initiative (ADNI) database. Neural Regeneration Research, 2018, 13 (6): 962-970.