Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.070912
ISSN 2371-8838 Vol. 7 Num. 9

Prediction of Alzheimer's Disease Based on Random Forest Model

Anran Lei[1,*], Jin Wang[2], Shicheng Zhou[2]

[1]School of Electrical Engineering and Automation, Hubei Normal University, Huangshi, 435002, China
[2]School of Computer and Information Engineering, Hubei Normal University, Huangshi, 435002, China
[*]Corresponding author: 15807210442@139.com

Keywords: Alzheimer's Disease, Logistic Regression Model, Random Forest Model, REF, SMOTE Algorithm

Abstract: Alzheimer's disease is a syndrome characterized by acquired cognitive impairment, leading to significant declines in daily life, learning, work, and social functioning. It has a profound impact on the lives of elderly people, making early detection and treatment of Alzheimer's disease an urgent issue. This paper collects relevant data from patients with Alzheimer's disease in a certain hospital, explores the data using histograms, density probability graphs, box plots, and correlation coefficient heat maps after preprocessing. Then it compares the performance of logistic regression classification models, random forest classification models, and REF-random forest models in predicting the accuracy of Alzheimer's disease categories. The results show that the REF-random forest model achieves the highest prediction accuracy. Finally, this paper uses the SMOTE algorithm to process the data and further improve the accuracy of the model. The optimized REF-random forest model has achieved outstanding results in all indicators.

1. Introduction

Alzheimer's disease is a syndrome characterized by acquired cognitive impairment, leading to significant declines in daily life, learning ability, work ability, and social interaction skills. The cognitive impairment in patients involves memory, learning, orientation, understanding, judgment, calculation, language, visual-spatial functioning, analysis, and problem-solving abilities. It often occurs with mental, behavioral, and personality abnormalities at a certain stage of the disease course. Its etiology may be related to genetic factors, unhealthy lifestyles, cerebrovascular disease, hyperlipidemia, and other factors.[1] With the increasing aging of our population, Alzheimer's disease has seriously affected the lives of the elderly. Therefore, it is urgent for us to detect and treat Alzheimer's disease patients as soon as possible. Before this paper, Fan Yu [2]used multiple machine learning models for prediction but did not optimize or improve them. This paper will use a double-optimized random forest model to predict them.

2. Data Preprocessing and Exploration

The data provided by the ANDI database (Alzheimer's Discase Neuroimaging Initiative) consists of participants aged 55-90 years old who are able to provide independent functional assessments and have been screened for specific psychoactive drugs. The data includes RID, EXAMDATE, DX_bl, AGE, PTGENDER, PTEDUCAT, PTETHCAT, and more. Among these variables, DX_bl is a classification variable for Alzheimer's disease, including categories such as CN (Normal Cognitive Decline), AD (Alzheimer's Disease), LMCI (Mild Cognitive Impairment), SMC (Senile Mild Cognitive Decline), and EMCI (Early Mild Cognitive Impairment).

Out of the selected data, there are 1738 cases of AD, 4850 cases of CN, 2968 cases of EMCI, 5236 cases of LMCI, and 1416 cases of SMC. Below is an explanation of each category:

CN: Normal Cognitive Decline. Participants in the CN group are the control subjects in this study. They do not exhibit signs of depression, mild cognitive impairment, or dementia.

SMC: Senile Mild Cognitive Decline. This category aims to bridge the gap between healthy elderly controls and MCI individuals.

MCI: Mild Cognitive Impairment. MCI participants maintain daily activities and show no significant impairment in other cognitive domains without evidence of dementia. The MCI level is determined using the Wechsler Memory Scale Logical Memory II (either early or late).

EMCI: Early Mild Cognitive Impairment.

LMCI: Late Mild Cognitive Impairment.

AD: Alzheimer's Disease. [3]

Some partial data is shown in Table 1.

Table 1: Shows some partial data.

ORIGPROT	SITE	VISCODE	EXAMDATE	DX_bl	AGE	PTGENDER	PTEDUCAT
ADNI1	11	bl	2005/9/8	CN	74.3	Male	16
ADNI1	11	bl	2005/9/12	AD	81.3	Male	18
ADNI1	11	m06	2006/3/13	AD	81.3	Male	18
ADNI1	11	m12	2006/9/12	AD	81.3	Male	18
ADNI1	11	m24	2007/9/12	AD	81.3	Male	18
ADNI1	22	bl	2005/11/8	LMCI	67.5	Male	10
ADNI1	22	m06	2006/5/2	LMCI	67.5	Male	10
ADNI1	22	m12	2006/11/14	LMCI	67.5	Male	10

2.1 Data Preprocessing

In this study, the proportion of missing values in each variable was first examined. Variables with missing values greater than 0.5 were removed as they had no research significance. The remaining missing values were then processed by replacing numerical values with their mean and storing non-numerical values in an "append" column. For variables with missing values greater than 0.1, if it was a missing value, it was labeled as 0; otherwise, it was labeled as 1. Next, the variables CN, AD, LMCI, SMC, and EMCI in the categorical variable were assigned numbers 0 to 4 for further analysis. Data processing was completed.

2.2 Data Exploration

To explore the processed data, this study plotted histograms, density probability graphs, and box plots to visualize the distribution of the data. Additionally, using Spearman's coefficient, a heatmap of correlation coefficients was created to examine the correlation between variables. Due to the limited space in this paper, only some data is presented.

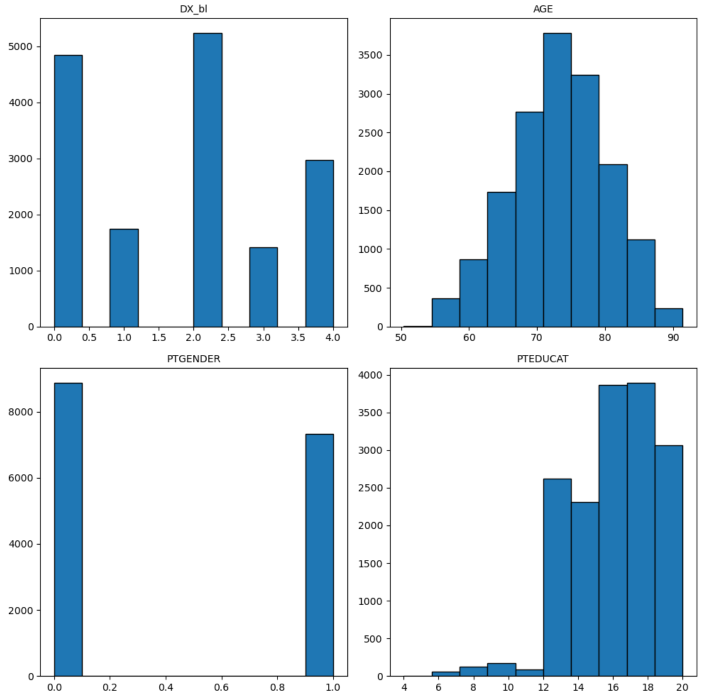

Figure 1: Histogram.

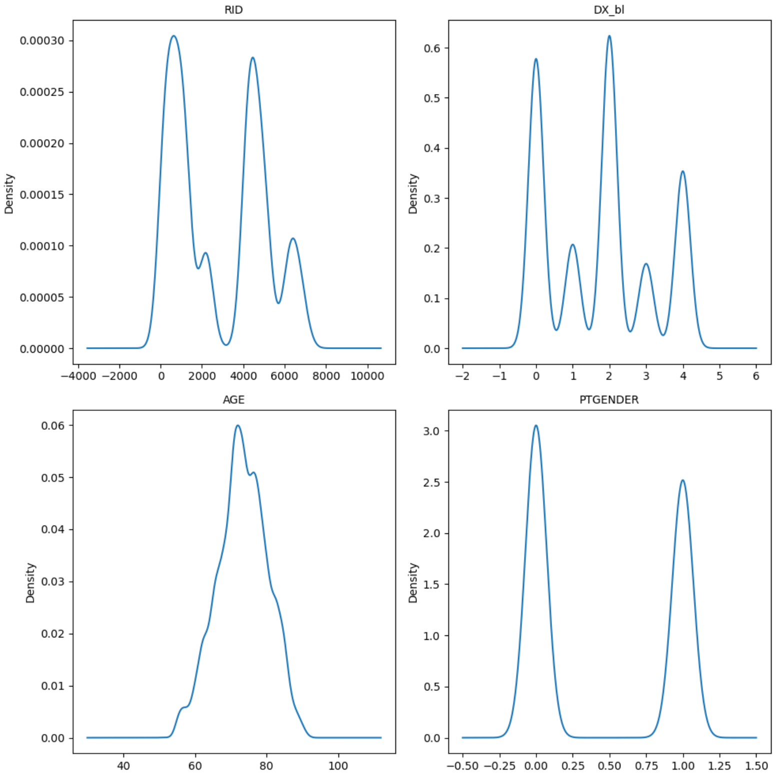

Figure 2: Probability density graph.

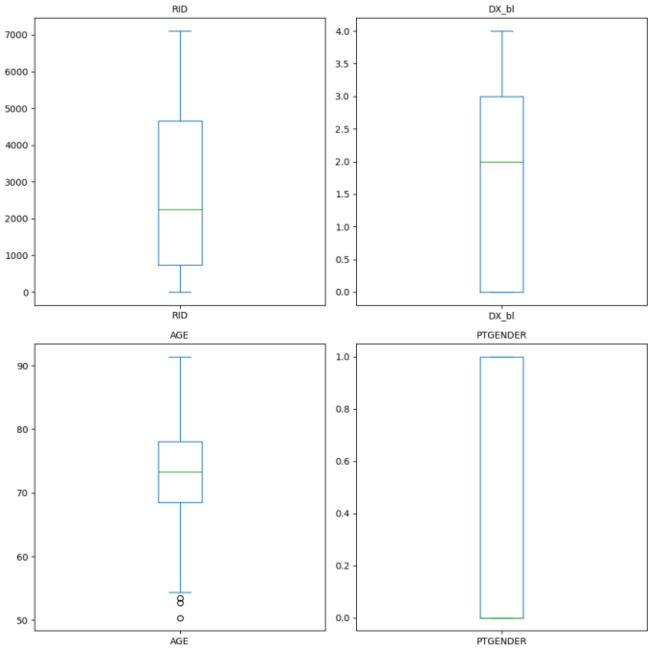

Figure 3: Box plot.

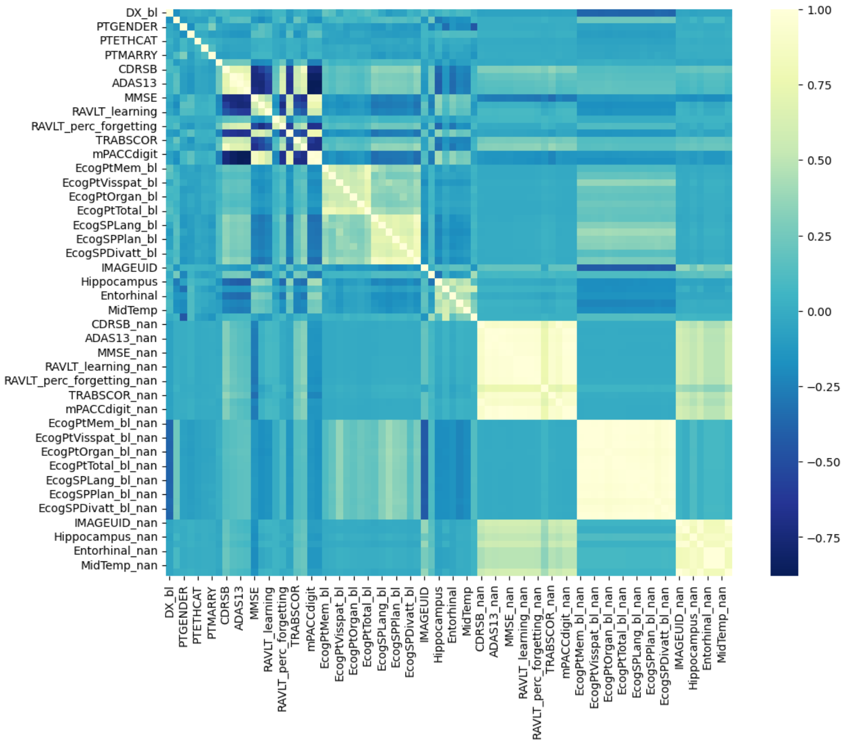

Figure 4: Heatmap of correlation coefficients.

Through the histogram (Figure.1) and density graph (Figure.2), we can observe that the age follows a normal distribution, indicating that the selected data have a certain generalizability. However, other labels do not conform to the normal distribution. The box plot (Figure.3) reveals that there are some outliers in some data, but since this sample size is relatively large, they can be ignored. By observing the heatmap (Figure.4), we can see that some data have strong correlations, but most data do not show significant correlations.

3. Model Selection

3.1 Introduction to the Base Models

In this study, two base models were chosen for multi-class classification: Logistic Regression and Random Forest. The Logistic Regression model uses a multi-class logistic regression approach, treating each class as a binary problem compared to the remaining classes. For N classes, N-1 binary classifications are performed, resulting in N-1 binary models. The probability of each binary classification is calculated, and the class with the highest probability is used as the predicted result for the new sample.

The Random Forest model is an ensemble learning method that builds multiple decision trees from a random sample of the original dataset. In this study, we used Bootstrapping to randomly select n training samples from the original dataset for each iteration. A total of k iterations are performed, resulting in k training sets that are mutually independent. Each training set is used to train a single decision tree model. Finally, for the multi-class problem, these k decision tree models are used for classification by majority voting [4-6].

To address issues with irrelevant or lowly correlated variables affecting model accuracy, we employed a REF model based on the Random Forest model to select the most relevant indicators. Nine indicators with the highest correlation were selected and used as input features for the Random Forest model for prediction.

3.2 Model Solution

In this study, the processed data was used to train both Logistic Regression and Random Forest models. The results are presented in Tables 2 and 3.

Table 2: Logistic Regression Classification Evaluation Report.

	precision	recall	f1-score	support
0	0.19	0.32	0.24	900
1	0.00	0.00	0.00	2
2	0.80	0.37	0.51	3356
3	0.16	0.28	0.20	244
4	0.16	0.38	0.22	361
accuracy			0.36	4863
macro avg	0.26	0.27	0.23	4863
weighted avg	0.60	0.36	0.42	4863

Table 3: Random Forest Classification Evaluation Report.

	precision	recall	f1-score	support
0	0.91	0.90	0.90	1506
1	0.70	0.86	0.77	410
2	0.88	0.86	0.87	1609
3	0.75	0.90	0.82	359
4	0.95	0.85	0.90	979
accuracy			0.87	4863
macro avg	0.84	0.87	0.85	4863
weighted avg	0.88	0.87	0.87	4863

REF model solution results:

Feature names: RID, AGE, EXAMDATE, EcogSPPlan_bl, LDELTOTAL, EcogPtPlan_bl, EcogPtOrgan_bl,TRABSCOR,EcogSPOrgan_bl.

Considering that the EXAMDATE variable is a date variable, we discard it and keep the eight variables except for it. Substituting them into the random forest model, the results are shown in Table 4.

Table 4: Random Forest Classification Evaluation Report after Ref Optimization

	precision	recall	f1-score	support
0	0.91	0.91	0.91	1408
1	0.66	0.88	0.76	380
2	0.91	0.85	0.88	1683
3	0.94	0.87	0.90	464
4	0.95	0.97	0.96	856
accuracy			0.89	4863
macro avg	0.87	0.90	0.88	4863
weighted avg	0.90	0.89	0.89	4863

Conclusion: From Table 2, we can observe that the accuracy of the five categories predicted by logistic regression is low, with the exception of category 2. After weighted averaging, the accuracy is only 60%, and the recall rate is also relatively low. Therefore, this paper abandons the use of this model for prediction. From Table 3, we can find that the random forest model has a high accuracy in predicting the five categories. The weighted average accuracy reaches 88%. After using the REF model to optimize the random forest model, the accuracy of category 1 has slightly decreased, while the accuracy of other categories has improved. Specifically, the accuracy of category 3 has increased by 19%, and the weighted average accuracy has also increased by 2%. Moreover, the recall rate has also improved by 2% compared to before optimization. Therefore, this paper chooses the REF-optimized random forest model for prediction.

4. An optimized prediction model based on the SMOTE algorithm

In order to solve the imbalanced class problem, this paper uses the SMOTE algorithm for synthetic data synthesis. It combines over-sampling of minority classes and under-sampling of majority classes to process variables.

Table 5: Classification evaluation report of REF-Random Forest after optimization using SMOTE algorithm.

	precision	recall	f1-score	support
0	1	1	1	1488
1	1	1	1	506
2	1	1	1	1565
3	1	1	1	431
4	1	1	1	873
accuracy			1	4863
macro avg	1	1	1	4863
weighted avg	1	1	1	4863

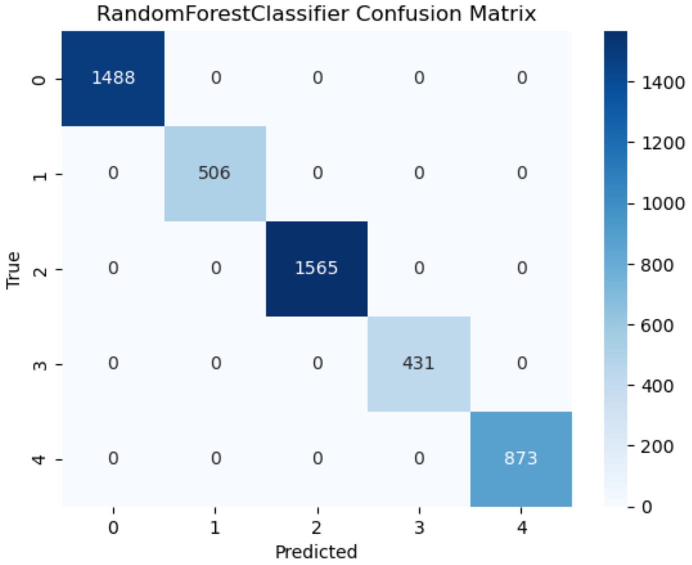

Figure 5: Confusion matrix of REF-Random Forest after optimization with SMOTE model.

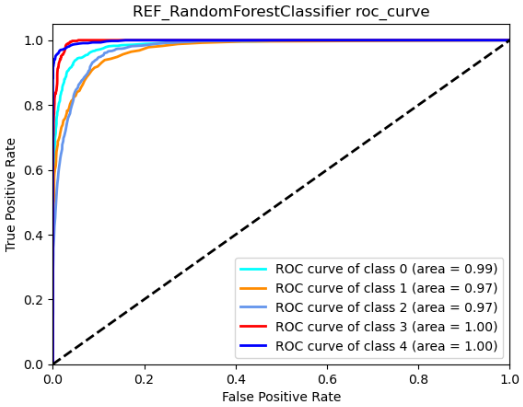

Figure 6: ROC curve of REF-Random Forest model.

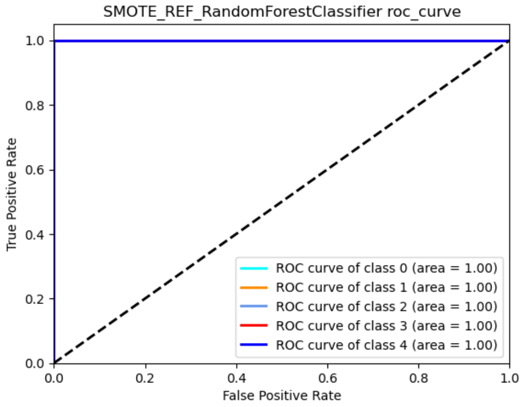

Figure 7: ROC curve of SMOTE-REF-Random Forest model.

From Table 5 and Figure 5, we can observe that the accuracy and recall rate of the REF-Random Forest model after being processed by the SMOTE algorithm have increased from approximately 90% before to 100%. It accurately classifies the data, indicating that this optimization was very successful. We then plotted the ROC curve for each category before and after optimization, as shown in Figures

6 and 7. It can be seen that the results were already good before optimization, but after optimization, they reached perfection.

5. Conclusion

This paper concludes that factors significantly affecting Alzheimer's disease include AGE, EcogSPPlan_bl, LDELTOTAL, EcogPtPlan_bl, EcogPtOrgan_bl, TRABSCOR, and EcogSPOrgan_bl. By comparing the accuracy and recall rate of the REF model with those of logistic regression and random forest models, this study finds that the REF-random forest model has the best prediction effect. Finally, after using the SMOTE algorithm to optimize the REF-random forest model, both recall rate and accuracy have reached 100%. As shown in Figure 5, the confusion matrix also indicates that the model accurately classifies each category. Hospital examination reports can be imported into this model to predict which category a patient belongs to (CN, AD, LMCI, SMC, EMCI), thus enabling early detection and timely treatment. This will delay disease progression, improve quality of life, and reduce the burden on both individuals and society as a whole.

References

[1] Long Yuanxian, Tang Yumeng, and Tang Lijun. A Meta-Analysis of Main Influencing Factors for Alzheimer's Disease in China [J]. Chinese Journal of Preventive Medicine, 2013, 14(01): 59-63.

[2] Fan Yu, Chen Tingting, and Chen Gang. Construction of Alzheimer's Disease Absent-Home Atrophy-Related Prediction Models Using Multiple Machine Learning Models [J/OL]. Bioinformatics: 1-14 [2023-08-31].

[3] Mao Nannan. Classification Research on Alzheimer's Disease Based on Multimodal Feature Fusion [D]. Dalian Maritime University. 2021.

[4] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com.

[5] Chang Jucai, Qi Pengfei, and Chen Xiao. Multi-condition Rock Hardness Recognition for Roadheader Based on Feature Selection and Random Forest [J]. Journal of China Coal Society, 2023, 48(2): 1070-1084.

[6] Tao Jintao, Zhang Nannan, Chang Jinyu, et al. Three-dimensional Ore Deposit Prediction Research Based on Logistic Regression for the Honghai Ore Deposit in the East Tianshan Mountains [J]. Xinjiang Geology, 2022 (040-001).