Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/49/20241045

Alzheimer's disease intelligent detection combining XGBOOST and NARX

Peng Wei

School of International College of Transportation, Chang'an University, Xi'an, China

peng.wei@ucdconnect.ie

Abstract. Due to the current situation of mental health illness, which causing a huge impact on the society. In this paper, an attempt has been made to analyses and predict the data from ANDI using single and composite algorithms. This paper used chi-square test, Spearman's correlation coefficient and maximum mutual information number, cost-sensitive learning, SMOTE, ADASYN, SMOTE+ENN, SMOTE+TOMEK to investigations. Specifically, this paper adopted the random forest to fill the data, and besides, given the fact that the data shows the characteristics of imbalance, this paper identifies the method of SMOTE TOMEK integrated sampling, XGBOOST and Bayesian optimization scheme to give the best performance and the best classification was obtained by XGBOOST combined with SMOTE-TOMEK. Furthermore, this paper used the NARX network to track the changes generated by time-based indicators, providing another insight to refine the study of intelligent diagnosis of Alzheimer.

Keywords: Alzheimer's Disease, SMOTE-TOMEK, Integrated Algorithm, NARX Network.

1. Introduction

Alzheimer's Disease International published the World Alzheimer Report in 2018, which states that dementia has become a global crisis. Based on the data, approximately 50 million people had Alzheimer's disease globally in 2018, that number is expected to grow to 82 million by 2030 and to 0.152 billion by 2050. Alzheimer's disease accounts for more than half of the dementia population. A large number of people with AD miss out on diagnosis due to lack of intervention in the mid-stage, and treatment is ineffective in patients with advanced AD because the nerves have already been damaged. That's why scientists and doctors are eager to accurately detect and intervene in early-stage Alzheimer's patients.

Choi et. al. study predicting cognitive decline with deep learning of brain metabolism and amyloid imaging [1]. The network's design exclusively uses baseline PET analyses and healthy participants as the dataset of train [1-3]. In order to identify AD patients and healthy control participants, Lahmiri and Shmuel set out to determine the extent to which specific characteristics, such as the index of gyrification, the results of the AD cognitive test, and the breadth of the cortex, are useful [2,3]. A strong ensemble deep learning model based on stacked CNNs and bidirectional long short-term memory networks is proposed by El-Sappagh et al [3]. By employing this technique, Yuan et al. significantly improve upon the baseline of 75.0% set by the challenge's creators to obtain 89.6% accuracy in the DementiaBank dataset [4]. The novel AD multiclass classification system is proposed by Zhang et al. based on multimodal neuroimaging and embedding feature selection and fusion [5]. Jia aims to create a group of

DOI: 10.54254/2755-2721/49/20241045

indicators that might be used to identify Alzheimer's disease (AD) at an early stage and to build risk prediction models that can predict the probability of AD occurrence and progression among China's older population during a period of five years [6]. A cohort of 2000 all-sex and all-ethnic individuals aged 60 and over who have a permanent address in the Tianma neighborhood, SheMountain Town, Songjiang District, Shanghai as part of the Wang research [7]. CZAMANSKI-COHEN suggests using 4,000 self-figure drawings to create and test an application designed to distinguish between drawings of people with MCI, AD, and healthy controls (HC) [8]. To comprehensively classify all research on genetic associations related to AD, Bertram et al. created a freely accessible, regularly updated database (http://www.alzgene.org) [9].

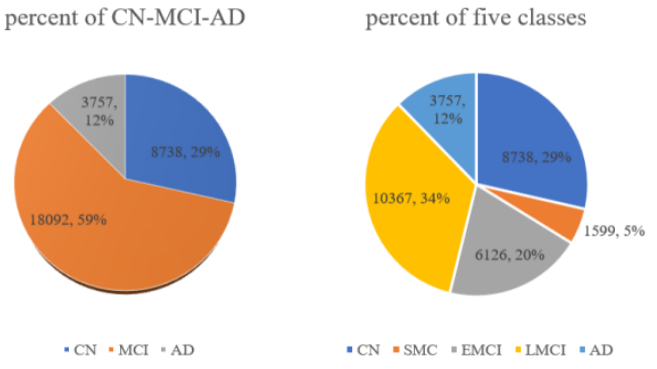
However, the current diagnostic solutions are time-consuming and current research focuses on data processing in the object dimension and lacks the temporal dimension to track patient progression. In this paper, this paper refers to the idea to extract data from the Department of Defense (DOD) ANDI. This paper focuses on the intelligent diagnosis of AD based on the structural and cognitive-behavioral characteristics of the brain in patients at different times.

2. Data

The attached data contain specific information characteristics of 5236 late mild cognitive impairment individuals (LMCI), 1416 patients with subjective memory complaint (SMC), 1738 clients with AD, 4850 cognitive normal elderly (CN) and 2968 early-stage mild cognitive impairment sufferers (EMCI) collected at different time points. This paper found that each piece of data has two identical expressions, which can be considered as two tests of the patient's physical indicators, split the results of the two feature tests and transformed one piece of data into two. By splitting the data, this paper gets the size of 32444*67. Some basic information of the individuals is shown below in Table 1 and Figure 1.

Table 1. Statistical analysis to detect data outliers.

	FDG	PIB	AV45	FBB	ABETA	TAU
count	32444	32444	32444	32444	32444	32444
mean	1.203170	1.722274	1.147617	1.236531	954.9938	217.1104
std	0.130655	0.229156	0.186401	0.135186	399.3969	97.99915
min	0.566989	1.095	0.809	0.8702	100	80
25%	1.13441	1.512175	1.017795	1.135210	676.9543	146.2589
50%	1.211314	1.7513	1.060403	1.204204	891.6455	168.4009
75%	1.283666	1.908338	1.2711	1.335189	1 128.0228	254.2948
max	1.77617	2.9275	2.6866	2.0088	2000	1400



(a) **(b)**

Figure 1. percent of classes.

DOI: 10.54254/2755-2721/49/20241045

3. Model assumptions and Model Construction

3.1. Model assumptions

This paper bases on the following assumptions to construct the detailed models:

- Only the characteristic indicator variables given in the Appendix are considered in the diagnosis of AD.
- The true validity of the annexed data is assumed.
- It is assumed that the missing values in the data have a certain pattern.
- Exclude errors recorded at the time of data set collection.
- Ignore the effect of outliers on the validity of the data.
- Changes in disease characteristics in question 4 are not confounded by remaining factors other than time. The missing values are tested and extracted and the dataset is complemented with random forest algorithm.

3.2. Data preprocessing

This paper uses Random Forest as the correction method for error value. Random Forest generated decimal data for rounding into integers, and then compared the complementary categorical data and the original data found that there is no anomaly (e.g., DX is CN but DXX is LMCI), proving that the prediction of the fitting effect is better. There are different ways of data balancing, in this paper this paper uses SMOTE, ADASYN, SMOTE-ENN and SMOTE-TOMEK respectively (See Table 2).

Table 2. CN, SMC, EMCI, LMCI, AD resampling results.

	Total number	CN	SMC	EMCI	LMCI	AD	Five Types of Ratios
Before Resampling	30587	8738	1599	6126	10367	3757	5:1:4: 6:2
SMOTE	36285	7257	7257	7257	7257	7257	1:1:1: 1:1
ADASYN	35927	6803	7508	7272	7257	7087	1:1:1: 1:1
SMOTE-ENN	26319	4836	6312	5049	4514	5608	1:1:1: 1:1
SMOTE-TOMEK	49409	9574	10072	9961	9687	10115	1:1:1: 1:1

3.3. Model building

For click-through rate and multi-classification prediction, GBDT is a frequently used technique, however it lacks superior computing efficiency and scalability [10]. However, the GBDT technique's computational efficiency is at least 20 times higher after the addition of GOSS and EFB, and the resulting algorithm is known as LightGBM [10]. The choice tree technique, which differs from the loss function in the standard tree, is used to enhance and optimize XGBOOST [11]. The XGBOOST algorithm is frequently downsampled by quantlie to make ensuring the samples are comparable to the original distribution and to discover the branching optima rapidly [11]. XGBOOST optimizes the direction categorization of the inputs of sparse matrices, which also contributes greatly to the optimization rate and computational speed [11]. With the step of scoring and best splitting the tree being retained, CatBoost enhances the first stage of decision tree building in comparison to the previous two algorithms (LightGBM and XGBOOST), which define correlation values in terms of leaves on a predetermined tree structure. Finally, an optimization that combats overfitting caused by gradient bias is obtained [12].

Firstly, different algorithms based on neural network, distance based, statistics based, and rule-based algorithms are applied in the training set, the algorithms used are single layer perceptron, KNN, and Naive Bayesian Model. The composite algorithms used are AdaBoost, XGBOOST, LightGBM, and CatBoost.

From the Table 3, it can be seen that XGBOOST, LightGBM and CatBoost outperform other algorithms in all indicators. So, these three methods are selected for further parameter optimization and

DOI: 10.54254/2755-2721/49/20241045

result comparison, which belong to tree model and Boosting algorithms, which mainly generate a series of weak classifier decision trees by training a subset of samples, and then obtain strong classifiers through the progressive combination of trees.

Table 3. Model classification performance evaluation metrics.

Algorithm	Accuracy	micro-Precision	micro-Recall	micro-F1
Perceptron	0.413534	0.413534	0.413534	0.413534
KNN	0.715049	0.715049	0.715049	0.715049
MNB	0.470197	0.470197	0.470197	0.470197
SVM	0.432930	0.432930	0.432930	0.432930
Logistic Regression	0.409938	0.409938	0.409938	0.409938
LightGBM	0.938760	0.938760	0.938760	0.938760
XGBoost	0.938433	0.938433	0.938433	0.938433
CatBoost	0.932549	0.932549	0.932549	0.932549
AdaBoost	0.627329	0.627329	0.627329	0.627329

3.4. Model optimization

Multiple structural connectivity networks can mathematically represent changes in Alzheimer's disease over time, according to research on the nonparametric Bayesian design of complex networks. As a result, Bayesian modeling techniques for different structured connection networks can be used to extract and demonstrate the correlation of indications of Alzheimer's disease progression [13]. The collection work of Bayesian optimized originates from the posterior distribution derived from the basic data set and is used to choose the next assessment point by maximizing the value. The cast function is maps from the input space, observation space, and hyper-parameter space to the real space. [14]. These techniques can enhance patient diagnosis and treatment while assisting the medical community in better comprehending and predicting the mechanics of Alzheimer's disease progression [13]. Figure 2 displays the Bayesian tuning diagram.

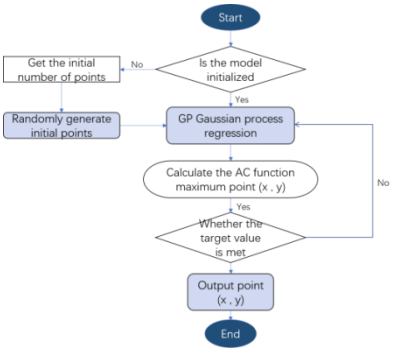


Figure 2. Bayesian Tuning Flow Chart.

This paper takes the sample data with RID=4, 35, 41, 58, 229, 1190, 2061, 4466, 4526, do the splitting and expansion of the data, and renumber and resample the data several times to avoid the error of prediction due to the difference of the sample ordinal number, until the regression R value of the training set is stable above 0.95, and not lower than 0.80 in validation and testing.

The NARX network was used, and the training algorithm was Levenberg-Marquardt. The mean square error and MSE were used for evaluation, and the data set was divided into 70:15:15, and the obtained mean square errors were 0.1615, 0.9561, and 0.8557, the R values were 0.9641, 0.8643, and 0.8150.

DOI: 10.54254/2755-2721/49/20241045

3.5. Model Testing

This paper saved the final results of analyzing the data characteristics and the correlation of AD, and the visualization is shown below in Figure 3.

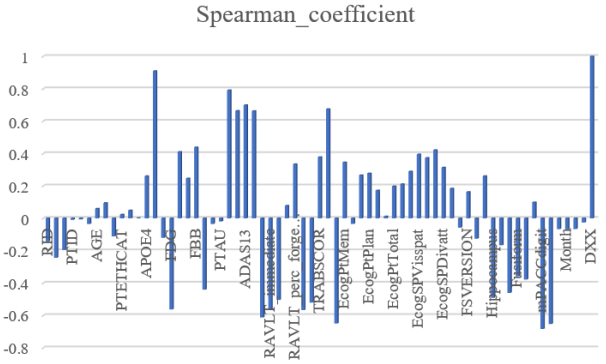


Figure 3. coefficient visualization.

Based on the results, this paper can conclude that AD has the highest correlation with the data feature CDRSB, higher correlation with the features ADAS11, ADAS13, ADASQ4, FAQ, MOCA, and mPACCdigit, and lower correlation with the features PTID, SITE, TAU, and PTAU.

Each step is fitted to the corresponding training set data, and finally the trained model with optimal parameters is applied to the test set by Bayesian optimization, and the Table 4 below displays the findings of the evaluation.

Table 4. Bayesian optimization model evaluation results.

Algorithm	Balancing Method	Accuracy	micro-Precision	micro-Recall	micro-F1
LightGBM	Cost-Sensitive Learning	0.937561	0.937561	0.937561	0.937561
	SMOTE	0.933638	0.933638	0.933638	0.933638
	ADASYN	0.932658	0.932658	0.932658	0.932658
	SMOTE-ENN	0.891468	0.891468	0.891468	0.891468
	SMOTE-TOMEK	0.982456	0.982456	0.982456	0.982456
XGBoost	Cost-Sensitive Learning	0.938433	0.938433	0.938433	0.938433
	SMOTE	0.934074	0.934074	0.934074	0.934074
	ADASYN	0.935382	0.935382	0.935382	0.935382
	SMOTE-ENN	0.888853	0.888853	0.888853	0.888853
	SMOTE-TOMEK	0.986815	0.986815	0.986815	0.986815
CatBoost	Cost-Sensitive Learning	0.933312	0.933312	0.933312	0.933312
	SMOTE	0.930151	0.930151	0.930151	0.930151
	ADASYN	0.930151	0.930151	0.930151	0.930151
	SMOTE-ENN	0.878174	0.878174	0.878174	0.878174
	SMOTE-TOMEK	0.967746	0.967746	0.967746	0.967746

Model selection: As seen from the table, for the five balancing methods, all three algorithms perform best on SMOTE-TOMEK, so this paper choose to train on the data after SMOTE + TOMEK balancing. Based on the three existing models, this paper fused the models twice by soft-voting (Voting-Soft) on the three models and soft-voting on LightGBM + XGBOOST and used the five acquired models to examine the effects of the model on the test set. This study found that XGBOOST performs the best across all criteria.

DOI: 10.54254/2755-2721/49/20241045

In summary, this paper established a balanced-five-category full-process optimal model as an intelligent diagnostic model for AD, which can be used as model input by data like the structural brain characteristics and cognitive-behavioral characteristics in the attachment, and our model will output intelligent diagnostic results for AD (one of the five categories CN, SMC, EMCI, LMCI, AD). For the first sub-question of question three, the dataset is new_dataset.csv with the label [CN, MCI, AD], for the second sub-question of question three, the dataset is the new data consisting of the DX field in new_dataset.csv as MCI with the label [SMC, EMCI, LMCI] (See Tables 5-9 and Figures 4-7).

Note: The final optimization results for SMC, EMCI, and LMCI are given above.

Table 5. CN, MCI, AD resampling results.

	Total number of samples	CN	MCI	AD	Three types of ratios
Before resampling	30227	8738	18092	3757	2:5:1
SMOTE	37992	12664	12664	12664	1:1:1
ADASYN	38780	12816	12664	13300	1:1:1
SMOTE-ENN	29104	9448	8259	11397	1:1:1
SMOTE-TOMEK	52230	17403	17284	17903	1:1:1

Table 6. CN, MCI, AD final optimization results Algorithm Balancing Method result.

Algorithm	Balancing Method	Accuracy	micro-Precision	micro-Recall	micro-F1
LightGBM	Cost-Sensitive Learning	0.942138	0.942138	0.942138	0.942138
	SMOTE	0.942247	0.942247	0.942247	0.942247
	ADASYN	0.941157	0.941157	0.941157	0.941157
	SMOTE-ENN	0.925466	0.925466	0.925466	0.925466
	SMOTE-TOMEK	0.975155	0.975155	0.975155	0.975155
XGBoost	Cost-Sensitive Learning	0.943010	0.943010	0.943010	0.943010
	SMOTE	0.941484	0.941484	0.941484	0.941484
	ADASYN	0.941266	0.941266	0.941266	0.941266
	SMOTE-ENN	0.925793	0.925793	0.925793	0.925793
	SMOTE-TOMEK	0.987360	0.987360	0.987360	0.987360
CatBoost	Cost-Sensitive Learning	0.939087	0.939087	0.939087	0.939087
	SMOTE	0.938651	0.938651	0.938651	0.938651
	ADASYN	0.937452	0.937452	0.937452	0.937452
	SMOTE-ENN	0.921543	0.921543	0.921543	0.921543
	SMOTE-TOMEK	0.969162	0.969162	0.969162	0.969162

Table 7. Model fusion effect.

Algorithm	Accuracy	micro-Precision	micro-Recall	micro-F1
LightGBM	0.975155	0.975155	0.975155	0.975155
XGBoost	0.98736	0.98736	0.98736	0.98736
CatBoost	0.969162	0.969162	0.969162	0.969162
3-soft	0.980277	0.980277	0.980277	0.980277
2-soft(LightGBM+ XGBoost)	0.983328	0.983328	0.983328	0.983328

DOI: 10.54254/2755-2721/49/20241045

Table 8. SMC, EMCI, LMCI resampling results.

	Total number of samples	SMC	EMCI	LMCI	Three types of ratios
Before resampling	18092	1599	6126	10367	1:3:5
SMOTE-TOMEK	30477	10308	10085	10084	1:1:1

Table 9. Model fusion effect.

Algorithm	Accuracy	micro-Precision	micro-Recall	micro-F1
LightGBM	0.99963154	0.99963154	0.99963154	0.99963154
XGBoost	0.99944731	0.99944731	0.99944731	0.99944731
CatBoost	0.99963154	0.99963154	0.99963154	0.99963154
3-soft	0.99944731	0.99944731	0.99944731	0.99944731
2-soft (LightGBM+XGBoost)	0.99944731	0.99944731	0.99944731	0.99944731

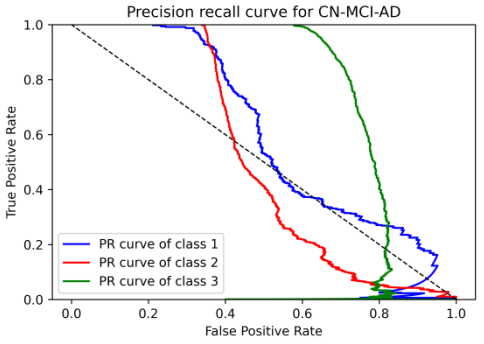


Figure 4. PR Curve for CN-MCI-AD.

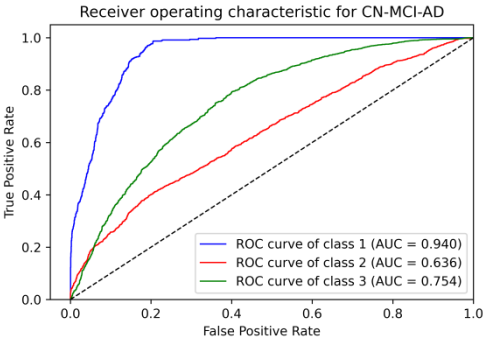


Figure 5. ROC curve for CN-MCI-AD.

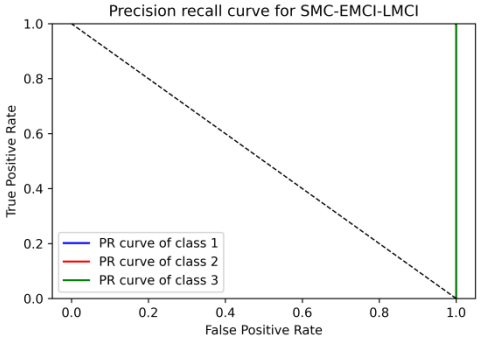


Figure 6. PR Curve for SMC-EMCI-LMCI.

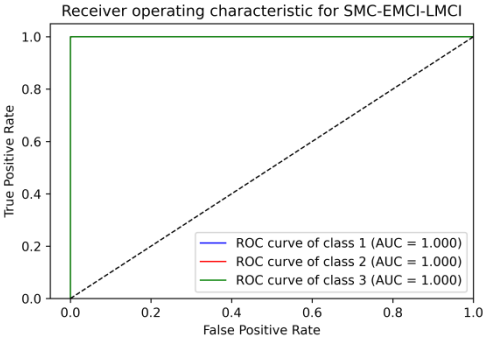


Figure 7. ROC curve for SMC-EMCI-LMCI.

The ROC curve of class I represents CN, class II ROC curve represents MCI, and class III ROC curve represents AD. The PR curves are more reflective of the classification performance when there is a large disparity in sample proportions. As can be seen from Figure 4, the more convex to the upper right the PR curve corresponding to AD is among the three types of PR curves, the better the effect is than the other two types. The AUC, which is used to reflect the validity of the detection method, is the

DOI: 10.54254/2755-2721/49/20241045

area under the section of the ROC curve. With the largest value in Figure 5 is 0.94 for the sign AUC, it indicates that CN categories has the best diagnostic efficacy compared with others.

The overall mrico-F1 index is higher implying the overall classification performance is better.

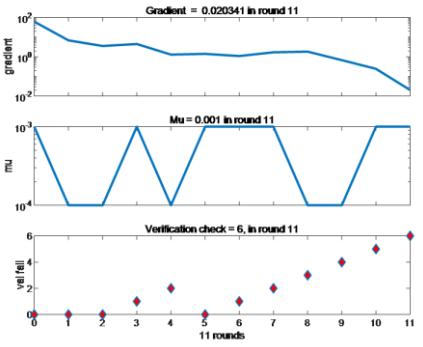


Figure 8. Diagram of the training phase.

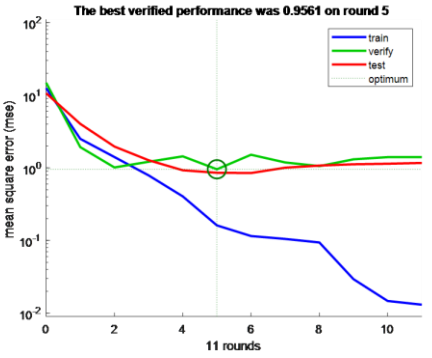


Figure 9. model Performance.

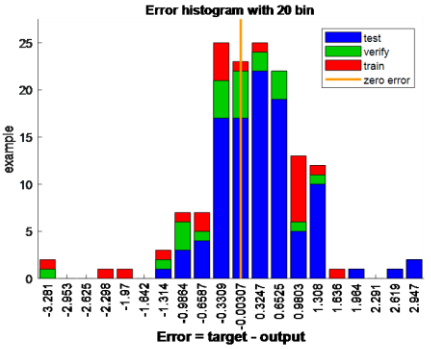


Figure 10. histogram error.

This paper used the NARX network as a time series to track the changes produced by time-based metrics for the age of specific patient characteristics by sampling the balanced data. From the above Figures 8-10, different kernel functions and cut ratios to the dataset are used to achieve regression R values stable above 0.95, 0.80 and 0.80, respectively. Therefore, this paper believes that the model this paper developed is better able to reveal the evolution of different categories of diseases over time (See Figures 11-12).

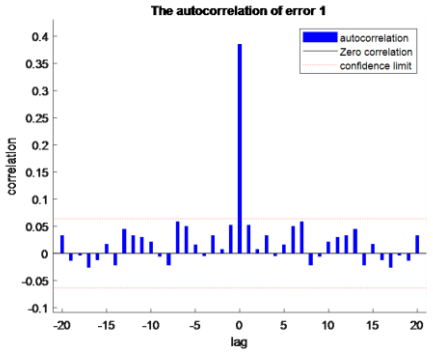


Figure 11. Auto correlogram.

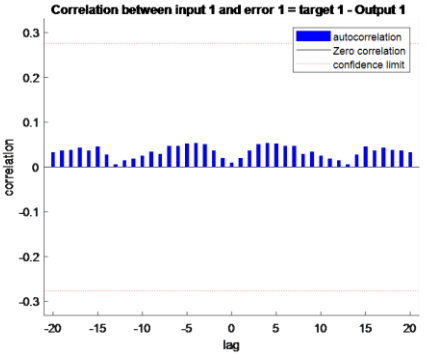


Figure 12. Input-error correlation.

DOI: 10.54254/2755-2721/49/20241045

4. Conclusions

In this paper, this paper focuses on the intelligent diagnosis of AD based on the structural and cognitive-behavioral characteristics of the brain in different categories of people. This paper added a more detailed five-category DXX (CN, SMC, EMCI, LMCI, AD) field to the original three-category label DX (CN, MCI, AD). Then, this paper performed data cleaning operations including coding, processing missing values, error values, abnormal values, and duplicate values. After comparing various methods of filling missing values, this paper built a random forest regression model to ensure the correct rate. Based on the preprocessing results, this paper used chi-square test, Spearman's correlation coefficient and maximum mutual information number, to explore the correlation between data characteristics and AD diagnosis. To design an intelligent diagnosis of AD, this paper developed a balanced-five category full-process optimal model. Training and test sets were randomly split by 70:30 (random_state=2022). This paper finally settled on the SMOTE TOMEK integrated sampling + XGBOOST + Bayesian optimization scheme solution model after comparing data balancing methods and classification algorithms. This paper also established a balancing-triple classification full-process optimal model to classify three major classes, as well as three subclasses to enhance the condition differentiation and intelligent diagnosis performance of AD. Furthermore, this paper used the NARX network as a time series to track the changes generated by time-based indicators of the age of specific patient characteristics by sampling balanced data to reveal the evolution of different categories of diseases over time, refining the study of intelligent diagnosis of AD from another perspective. This paper establishes an intelligent diagnostic model for AD through structural brain characteristics and cognitive-behavioral characteristics, which has better results in testing. The development of things is typically congruent with the utilization of statistics to foresee the future according to previous patterns. When analyzing the development trend, it is fairer to concentrate on the effects of seasonal and cyclical variations on particular time points [15]. Therefore, it can withstand some of the effects of random variables on the final results in this case. In conclusion, this paper verified the utility and validity of the model, which has a high reference value for the intelligent and accurate diagnosis of AD.

However, this paper ignores the factors that have an impact on AD other than the characteristics given by the dataset, which to a certain extent reduces its usefulness in real life.

References

[1] Choi, H., Jin, KH. (2018). Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. Behavioural Brain Research: An International Journal.

[2] Lahmiri, S., Shmuel, A. (2019). Performance of machine learning methods applied to structural mri and adas cognitive scores in diagnosing alzheimer's disease. Biomedical signal processing and control, 52(JUL.), 414-419.

[3] El-Sappagh, S., Abuhmed, T., Islam, S. M. R., & Kwak, K. S. (2020). Multimodal multitask deep learning model for alzheimer's disease progression detection based on time series data. Neurocomputing.

[4] Yuan, J., Bian, Y., Cai, X., Huang, J., & Church, K. (2020). Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease. Interspeech 2020.

[5] Zhang, Y. Q. P. (2021). Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion. Information Fusion, 66(1).

[6] Jia, J. (2020). Study on Body Fluid, Gene and Neuroimaging Biomarkers for Early Diagnosis of Alzheimer's Disease.

[7] Wang, G. (2022). Shanghai Cognitive Impairment Study of The Elderly Population: SheMountain Cohort.

[8] CZAMANSKI-COHEN J. (2023). Developing An Artificial Intelligence System to Detect Mild Cognitive Impairment and Alzheimer's Disease Dementia Through Self-Figure Drawing: An Innovative Approach.

DOI: 10.54254/2755-2721/49/20241045

[9] Bertram, L., Mcqueen, M. B., Mullin, K., Blacker, D., & Tanzi, R. E. (2007). Systematic meta-analyses of alzheimer disease genetic association studies: the alzgene database. Nature Genetics, 2(1), S23-S23.

[10] Meng, Q. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Neural Information Processing Systems. Curran Associates Inc.

[11] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA, 785-794.

[12] Dorogush, A. V., Ershov, V., & Gulin, A. (2018). Catboost: gradient boosting with categorical features support.

[13] Peterson, C. B., Osborne, N., Stingo, F. C., Bourgeat, P., & Vannucci, M. (2020). Bayesian modeling of multiple structural connectivity networks during the progression of alzheimer's disease. Biometrics.

[14] Wang, X., Ding, C. & Chen, T. et al. (2022) Research on the Application of Bayesian-Optimized XGBoost in Minor Faults in Coalfields, Mathematical Problems in Engineering, vol. 2022, 3409468, 13.

[15] Xu, M., Zhang, D-F. et al. (2018). A systematic integrated analysis of brain expression profiles reveals YAP1 and other prioritized hub genes as important upstream regulators in Alzheimer's disease. Alzheimer's & Dementia, 14: 215 - 229.