Frontiers in Bioscience Landmark

Front. Biosci. (Landmark Ed) **2022**; 27(1): 037
http://doi.org/10.31083/j.fbl2701037

Original Research

Annotating whole genome variants and constructing a multi-classifier based on samples of ADNI

Juan Zhou[1], Yangping Qiu[1], Xiangyu Liu[1], Ziruo Xie[1], Shanguo Lv[1], Yuanyuan Peng[1], Xiong Li[1,*]

[1]School of Software, East China Jiaotong University, 330013 Nanchang, Jiangxi, China
*Correspondence: lixiong@ecjtu.edu.cn; lx_hncs@163.com (Xiong Li)
Academic Editors: Leyi Wei
Submitted: 11 June 2021 Revised: 29 November 2021 Accepted: 21 December 2021 Published: 19 January 2022

Abstract

Introduction: Alzheimer's disease (AD) is the most common progressive neurodegenerative disorder in the elderly, which will eventually lead to dementia without an effective precaution and treatment. As a typical complex disease, the mechanism of AD's occurrence and development still lacks sufficient understanding. **Research design and methods**: In this study, we aim to directly analyze the relationship between DNA variants and phenotypes based on the whole genome sequencing data. Firstly, to enhance the biological meanings of our study, we annotate the deleterious variants and mapped them to nearest protein coding genes. Then, to eliminate the redundant features and reduce the burden of downstream analysis, a multi-objective evaluation strategy based on entropy theory is applied for ranking all candidate genes. Finally, we use multi-classifier XGBoost for classifying unbalanced data composed with 46 AD samples, 483 mild cognitive impairment (MCI) samples and 279 cognitive normal (CN) samples. **Results**: The experimental results on real whole genome sequencing data from Alzheimer's Disease Neuroimaging Initiative (ADNI) show that our method not only has satisfactory classification performance but also finds significance correlation between AD and *RIN3*, a known susceptibility gene of AD. In addition, pathway enrichment analysis was carried out using the top 20 feature genes, and three pathways were confirmed to be significantly related to the formation of AD. **Conclusions**: From the experimental results, we demonstrated that the efficacy of our proposed method has practical significance.

Keywords: Unbalanced data; Multi-class classification; Multi-objective optimization

1. Introduction

AD is a neurodegenerative disease with a high incidence and possesses enormous threaten to elderly people, leading cause of memory loss and even dementia [1]. Early accurate diagnosis and effective preventive measures can delay the development of the disease. The diagnostic techniques of AD, such as cognitive testing [2], neuroimaging [3,4], biomarker detection [5–7] and genetic variants detection [8,9] and so on, have different advantages and disadvantages and all of them are constantly developing. Researchers believe that identify the causal genetic variations and understanding their underlying molecular mechanisms not only play important role in early diagnosis, but also may help for designing innovative medicine for the AD.

Genome-wide association studies (GWAS) as a data-driven strategy have been applied to locate genetic variants contributed to the risk of complex diseases. So far, thousands of risk sites have been reported to be associated with complex diseases and traits. The GWAS Catalog collects the findings of all high-quality published GWAS literatures, which provides prior knowledge for investigators [10–13]. However, GWAS faces two enormous challenges, namely reproducibility and heritability. For example, factors, such as sample scale, study cohorts, sequencing technology, statistic techniques and so on, can significantly influence the results of GWAS. Therefore, different studies may come to different conclusions, or even completely conflicting conclusions about genetic variants involvement in the complex disease onset and progression, which result in lack of reproducibility. Another challenge is that the uncovered susceptible variants only account for a limited proportion of the heritability for each complex disease [14–19]. For example, multiple common and rare variants have been reported to be associated with AD [20]. Although the APOEε4 allele consistently reproduced in lots of studies [21], only small proportion of AD patients hold the APOEε4 allele [22], namely low heritability. It means that there are other genetic variants with marginal effect also contributing to the risk of AD. Consequently, considering the epistatic interaction between genetic variants instead of univariate analysis may enhance the heritability of AD and be able to identify unknown risk variants.

To accurately identify the risk loci of AD, imaging genetics studies use neuroimaging endophenotypes to fill the gap between DNA variants and AD [23]. With using this potential strategy, FRMD6 was firstly considering to be associated with AD by leading to hippocampal atrophy [24]. Although this kind of strategy has achieved some success, it has two main limitations. One is that although the iden-

tified risk site is related to regions of interest (ROI), it does not mean that it is related to AD. The other is that the association analysis between multiple loci and multiple ROIs across the whole genome will greatly increase the computational burden.

In this study, we conduct a whole genome analysis for ADNI based on multi-classifier XGBoost [25–28], aiming to design an effective diagnostic method only with genome information and identify potential risk loci. The merits of our work is reflected in three points: firstly, in order to improve the interpretability of the study and reduce false positives, VEP [29] mutation annotation software is used to screen risk SNPs from the whole genome, and secondly, our method directly predicts the outcomes only with genome variants information, which avoids the defect that the variation identified by imaging genetics method is not strongly associated with individual phenotype; Finally, with using XGBoost, our method can accurately recognize the MCI samples and our method further validates that the *RIN3* has the strongest correlation to AD and pathway enrichment analysis results with the top 20 genes show that three pathways (Pyruvate metabolism, Glycine, serine and threonine metabolism and ABC transporters) are significant (p-value < 0.05). All these three pathways have been reported to be associated with AD [30–32].

2. Methods

2.1 Variants annotation based on VEP

Whole genome sequencing always derives millions of SNPs which can be partitioned into categories such as missense variants, synonymous variants, driver variants, passenger variants and so on. It means that not all of them are deleterious to gene function, let alone pathogenicity. The VEP tool can determine which genetic variants are deleterious with using SIFT scores [33]. Based on the sequence homology and physical properties of amino acids, SIFT evaluates the effect of each amino acid substitution on protein function.

In this study, we download the whole genome sequencing VCF files from the ADNI database (adni.loni.usc.edu) [34], which contains 809 samples and each sample holds over 388 million SNPs by the Ilumina Omni 2.5 M BeadChip. To enhance the biological meanings and reduce the computational burden for downstream analysis, we use VEP to filter variants (IMPACT is HIGH or IMPACT is MODERATE and SIFT <0.05 and BIOTYPE is protein_coding). After that, each sample approximately carries 1800 deleterious SNPs. Of note the SNPs contained in these samples are not exactly the same. Then, for each single sample, we encode gene corresponding to each deleterious SNP as '1' when IMPACT is MODERATE and as '2' when IMPACT is HIGH and the rest are '0'. Consequently, a matrix derived contains 809 samples and 16163 genes.

2.2 Filtration based on multi-objective criterion

16163 genes derive a huge combinatorial space, resulting in enormous challenge for optimization. Therefore, a reasonable way is to filter 16163 genes at the beginning. Since our purpose is to identify which genes are pathogenic for AD, intuitively we can evaluate each gene according its correlation to phenotypes. In this study, the CN samples are labeled as 0, the MCI samples are labeled as 1, and the AD samples are labeled as 2. Multi-objective optimization technique has been wildly used in various area [35,36]. To enhance the robustness of our method, a multi-objective technique is proposed to filter out redundant and irrelevant genes. Here, we separately design two complementary objectives, which results in removing the genes that are highly correlated to other genes and have low correlations to phenotypes.

2.2.1 Objective 1

The uncertainty of the phenotype Y can be quantified by Shannon entropy H(Y) as Eqn. 1.

$$\mathrm{H(Y)} = -\sum_{\mathrm{i=0}}^{2} \mathrm{P(yi)} \log_2 \mathrm{P(yi)} \tag{1}$$

where $\mathrm{p(y_0)}$ denotes the possibility of CN and $\mathrm{p(y_1)}$ represents the possibility of MCI and $\mathrm{p(y_2)}$ is the possibility of AD in cohorts.

Joint entropy of a single gene and phenotype Y can be defined as Eqn. 2. Let X be a gene with three kinds of values (0, 1 and 2) as mentioned above.

$$\mathrm{H(Y,X)} = -\sum_{\mathrm{x=0}}^{2}\sum_{\mathrm{y=0}}^{2} \mathrm{P(x,y)} \log_2 \mathrm{P(x,y)} \tag{2}$$

Then, we use mutual entropy I(Y|X) to evaluate the information contribution of a single gene to the phenotype Y (or vice versa) as Eqn. 3. The larger its value, the greater the contribution of the gene to the phenotype.

$$\mathrm{I(Y \mid X)} = \mathrm{H(Y)} + \mathrm{H(X)} - \mathrm{H(Y,X)} \tag{3}$$

2.2.2 Objective 2

We believe that a pathogenic gene shows significant difference across groups and should relatively consistent in MCI and AD samples. For objective 2, two components are designed for evaluating each gene. The first component uses entropy H(X) to select genes with consistent pattern (CP) in MCI and AD samples as Eqn. 4.

$$\mathrm{CP(X)} = \mathrm{H_{AD}(X)} + \mathrm{H_{MCI}(X)} \tag{4}$$

where $H_{AD}(X)$ and $H_{MCI}(X)$ are the entropy of gene X in AD and MCI samples, respectively. The smaller CP(X)

value, the more consistent pattern in cases samples.

In addition, the significant difference (SD) across groups of the pathogenic gene can be defined by their normalized frequencies as defined in Eqn. 5. Note that the gene value '0' means its function is normal, '1' means its function moderate damaged and '2' denotes highly damaged.

$$\mathrm{SD(X)} = \frac{|\mathrm{X_0}|}{|\mathrm{CN}|} + \frac{|\mathrm{X_1}|}{|\mathrm{MCI}|} + \frac{|\mathrm{X_2}|}{|\mathrm{AD}|} \tag{5}$$

where $\mathrm{X_0}$ denotes the frequency of gene value '0' in CN samples and $|\mathrm{CN}|$ is the number of CN samples, $\mathrm{X_1}$ denotes the frequency of gene value '1' in MCI samples and $|\mathrm{MCI}|$ is the number of MCI samples and $\mathrm{X_2}$ denotes the frequency of gene value '2' in AD samples and $|\mathrm{AD}|$ is the number of AD samples. Therefore, Eqn. 6 can evaluate each gene, and the larger Score, the more likely the gene is pathogenic.

$$\mathrm{Score} = \mathrm{SD(X)} - \mathrm{CP(X)} \tag{6}$$

2.3 Multi-class classification based on XGBoost

In this study, we download 809 VCF files from ADNI. Out of 809 samples, 808 samples include 279 CN samples, 483 MCI samples and 46 AD samples and the remaining one's disease state cannot be addressed, so that it was discarded.

XGBoost implements parallel tree boosting technique in a portable and efficient way [25] and it provides various packages such as R, Python, Ruby and so on. XGBoost help researchers to solve classification problem in an easy-use, friendly way. In this study, we use the R package of XGBoost. Before running XGBoost, several parameters should be addressed according to specific task. In Table 1, the important parameters are listed.

Table 1. The parameters of XGBoost.

Parameter	Value
booster	gbtree
eta	0.03
max_depth	6
gamma	3
objective	multi:softprob
subsample	1
num_class	3

3. Materials and measures

Data used in our study were downloaded from the ADNI database. The ADNI was initiated in 2003 and was built in three phases: ADNI-1, ADNI-2 and ADNI-GO. ADNI collects several types of data (e.g., clinical, genetic, MRI image, PET image and biospecimen) from volunteers (over 1500 adults recruited from the U.S. and Canada, ages 55 to 90), using strict protocols and procedures to ensure consistencies. The primary goals of ADNI are to find out an effective early diagnosis technique, understand the molecular mechanism of the onset and progression of MCI and AD and eventually design new medical and treatments. The ADNI More up-to-date information can be found on www.adni-info.org. In this study, we collected 809 VCF files recalled by CASAVA pipelines. Except only one sample has no information about disease state, other 808 samples can be partitioned into 483 MCI samples, 279 CN samples and 46 AD samples.

At the beginning, each sample carry about 388 million SNPs. After VEP filtration, each sample only holds variants which are deleterious. Then, the protein coding genes corresponding to deleterious SNPs are encoded as '1' when IMPACT is MODERATE and as '2' when IMPACT is HIGH and the rest are '0'. In total there are 16163 genes, so that it would cost an enormous computational burden for downstream analysis. After the filtration step based on multi-objective criterion, a matrix derived contains 808 samples and 866 genes, which n equals to 808 and p is 866 as shown in Fig. 1.

$$\begin{bmatrix} s_{11} & \cdots & \cdots & s_{1p} & y_1 \\ s_{21} & & & \vdots & \vdots \\ \vdots & & & \vdots & \vdots \\ \underbrace{s_{n1} \quad \cdots \quad \cdots \quad s_{np}}_{\boldsymbol{X}_{n^*p}} & & & & \underbrace{y_n}_{\boldsymbol{Y}_{n^*p}} \end{bmatrix}$$

Fig. 1. Data formulation.

In this study, we directly predict the sample state (AD, MCI and CN) only with gene function state (0, 1 and 2). Thus, it can be considered as a 3-class classification problem. However, the number of AD samples is significantly smaller than AD and CN samples, so that it is a typical unbalanced data.

$$Acc = \frac{TN + TP}{TN + FP + TP + FN} \tag{7}$$

$$\mathrm{Sen} = \frac{TP}{TP + FN} \tag{8}$$

Our primary goal is to design an early diagnosis

method of MCI and AD. Of note, the accurate identification of MCI and AD is more meaningful than identifying CN samples. Consequently, we use both accuracy (*Acc*) and sensitivity (*Sen*) measures to fairly evaluate our method as defined in Eqns. 7 and 8.

In addition, the cross-validation strategy, sometimes called out-of-sample testing, is to test the generalization ability of a classifier. Here, we apply 10-folds cross validation which divides all the 808 samples into 10 parts. Each iteration 9 equal parts are used as the training set, and the remaining part is used as the test set, repeating 10 times in total. The 10-folds cross validation give an insight on how the classifier will generalize to an unknown samples and flag overfitting and model bias issues. Note that, the proportion of different labels in each part is the same as that of the original data, that is, it is still unbalanced in each iterations.

4. Experimental results

To demonstrate the classification performance and biological meanings of our method, we analyze the experimental results in prediction accuracy, feature importance and pathway enrichment.

4.1 Prediction accuracy

To demonstrate the advantages of XGBoost on unbalanced data, we compare the results of XGBoost with Logistic Regression (LR). LR has been widely used in various research fields, such as bioinformatics, social science applications and so on. LR can be divided into two categories: binary LR and multinomial LR. The former is specific for the categorical response has only two possible outcomes, while the latter is suitable for three or more categories without ordering.

In Fig. 2, the horizontal axis represents 10 independent test sets of 10-folds cross validation. We can find that in almost all test sets, the XGBoost method is better than the LR method. The lowest accuracy of XGBoost is 0.68 and the highest accuracy of XGBoost is 0.79. On average, the accuracy of XGBoost is 0.73.

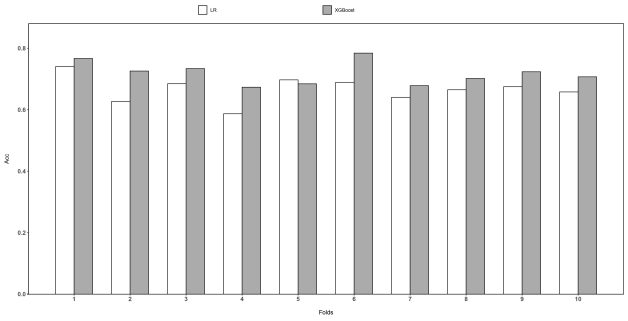

Fig. 2. The results of XGBoost and LR on accuracy.

In Fig. 3, XGBoost and LR have the completely same sensitivity on all test sets. For the 7th test set, all AD samples are not correctly identified by XGBoost or LR, while for the 1th test set, the sensitivity of both XGBoost and LR reach 0.8. It can be seen that too few AD samples cause the classifier to lack information to correctly distinguish AD samples from other samples.

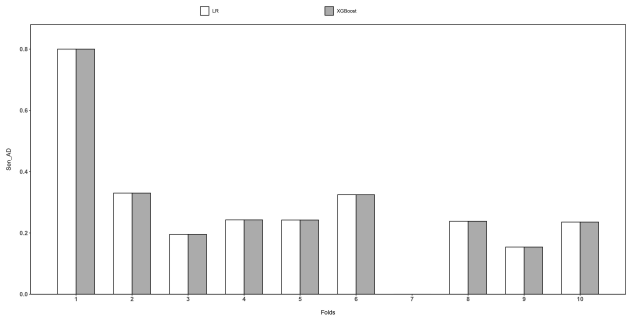

Fig. 3. The results of XGBoost and LR on sensitivity of AD.

In Fig. 4, the results show that XGBoost has better sensitivity of MCI than LR in all test sets. The lowest and highest results of XGBoost are 0.94 and 1.0, respectively. On average, the XGBoost's sensitivity of MCI is 0.97. These results demonstrate that XGBoost can accurately capture the characteristics of MCI samples, which would significantly for early diagnosis.

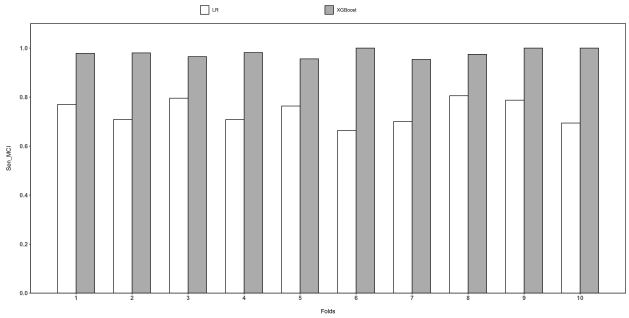

Fig. 4. The results of XGBoost and LR on sensitivity of MCI.

4.2 Feature importance

XGBoost classifiers samples in a decision tree ensembles manner. During the model training, each feature (gene) is evaluated by information gain which results in a rank of feature importance as shown in Fig. 5.

In Fig. 5, top 20 features, namely genes, are ranked according to their contributions in classification. Rab Interactor 3 (*RIN3*) is the most important feature than others. *RIN3*, a guanine nucleotide exchange factor (GEF) for the Rab5 small GTPase family, has been reported to be associated with both late onset AD (LOAD) and sporadic early onset AD (sEOAD) [37]. Shen *et al.* [38] validated that the upregulation of *RIN3* induces endosomal dysfunction in

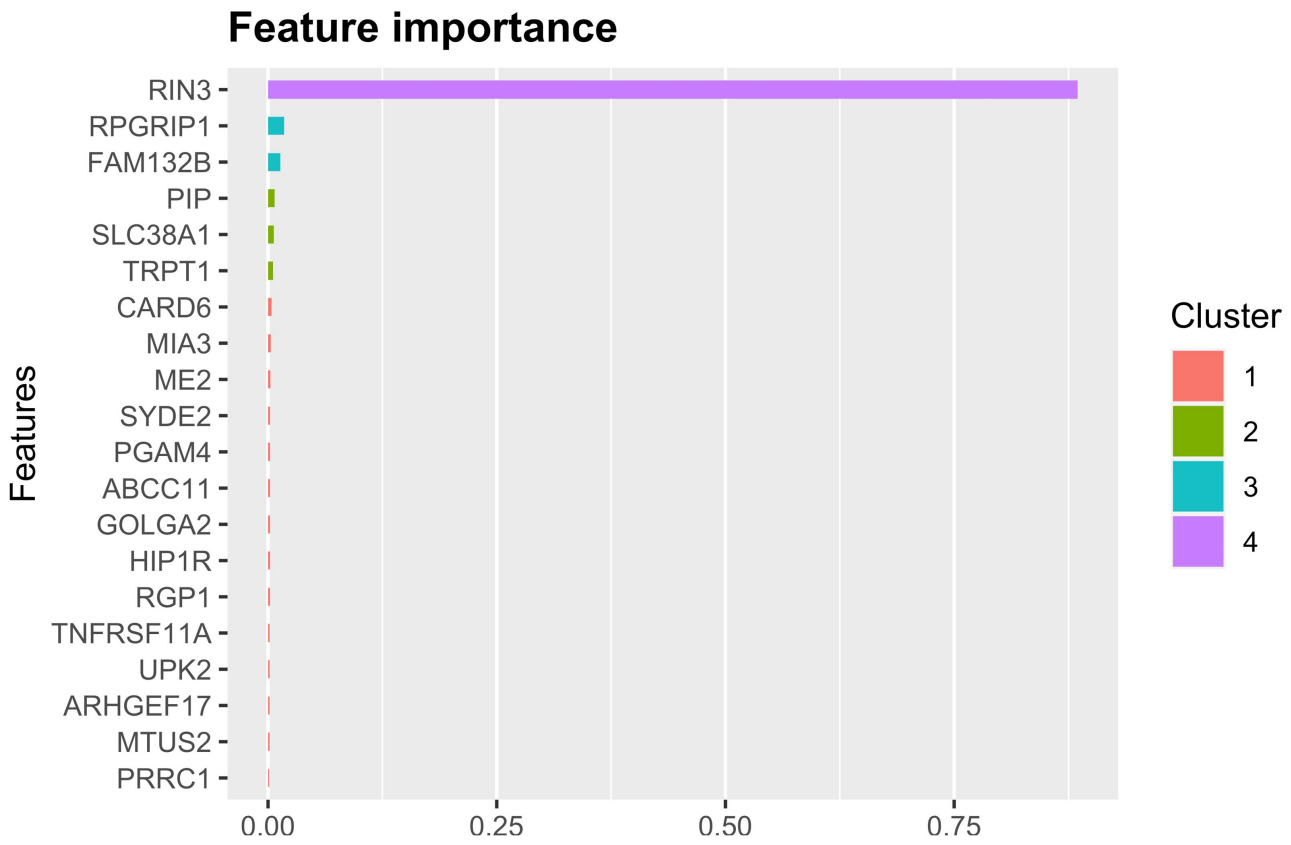

Fig. 5. The importance of top 20 features.

Table 2. Pathways with p-value < 0.05.

Name	p-value	Odds ratio	Combined score
Pyruvate metabolism	0.03830	25.64	83.65
Glycine, serine and threonine metabolism	0.03927	25.00	80.93
ABC transporters	0.04407	22.22	69.38

Alzheimer's disease, which provides a deeper understanding of how *RIN3* affects AD pathogenesis. In addition, the GWAS Catalog 2019 also shows that *RIN3* is a risk factor to the LOAD (p-value = 0.06682 and Odds Ratio = 15.63).

4.3 Pathway enrichment analysis

In further, we use the top 20 genes (*RIN3*, *RPGRIP1*, *FAM132B*, *PIP*, *SLC38A1*, *TRPT1*, *CARD6*, *MIA3*, *ME2*, *SYDE2*, *PGAM4*, *ABCC11*, *GOLGA2*, *HIP1R*, *RGP1*, *TNFRSF11A*, *UPK2*, *ARHGEF17*, *MTUS2* and *PRRC1*) to conduct a pathway enrichment analysis by KEGG 2019 Human in Enrichr [39].

In Table 2, the enrichment analysis results of three pathways with p-value < 0.05 are detailed. Moreover, all these three pathways have been reported to be associated with the development of AD. For example, Zilberter *et al.* [30] proved that in a mouse model of Alzheimer's disease Pyruvate pathway prevents the progression of age-dependent cognitive deficits without reducing amyloid and tau pathology. Oskouie *et al.* [31] investigated serum metabolic features in a mouse model of Alzheimer's disease and found that Glycine, serine and threonine metabolism pathway plays a key role in the onset and development of Alzheimer's disease. Pereira *et al.* [32] also deemed ABC transporters to be key factors in Alzheimer's disease.

5. Discussion

In this study, we introduced a framework for annotating whole genome variants and predicting Alzheimer's disease based on multi-classification: firstly, we use VEP to annotate the impact consequence of all single nucleotide polymorphism (SNP) and identify deleterious SNPs. After that, we map all these deleterious SNPs to nearest genes and design a multi-objective criterion to evaluate each protein coding gene. Finally, we apply a XGBoost for multiclass classification. From the experimental results on real whole genome sequencing data, we can find that our method achieves satisfactory classification performance. In addition, our method further validates that the *RIN3* has the strongest correlation to AD. Moreover, we use the top 20

genes to carry out pathway enrichment analysis and three pathways (Pyruvate metabolism, Glycine, serine and threonine metabolism and ABC transporters) are significant (p-value < 0.05). All these three pathways have been reported to be associated with AD.

Our method has several advantages: Firstly, unlike previous studies, our method predicts the sample outcomes according to the degree of damage to gene function. The function loss of gene annotated by VEP can be MODERATE or HIGH caused by SNPs, which enriches the biological meanings of our study; Secondly, our method directly predicts the outcomes only with genome variants information, which avoids the defect that the variation identified by imaging genetics method is not strongly associated with individual phenotype; Finally, our method filter the redundant and irrelevant SNPs by multi-objective evaluation, which significantly reduces the computational burden of classification model training. In addition, our method can accurately recognize the MCI samples, so that it would play important role in early diagnosis.

With using the feature importance evaluated by XGBoost, we find that the damage of *RIN3* gene function is significantly associated with the development of AD. This finding also has been proved by previous studies. Moreover, we examine the top 20 genes for pathway enrichment analysis. These genes enrich in Pyruvate pathway, Glycine, serine and threonine metabolism pathway and ABC transporters pathway with p-value < 0.05. Interestingly, these pathways have been confirmed to be related to the occurrence and development of AD. This result shows that the whole genome analysis in this paper is effective and XGBoost can identify the most important gene or even pathogenic gene.

The computational framework of our work has made some progress, but there are still deficiencies to be improved in future work. The limitations are as follows:

(1) Concerning the filter parameters (IMPACT is HIGH or IMPACT is MODERATE and SIFT <0.05 and BIOTYPE is protein_coding) of the pipeline, SNPs located in non-coding regions of relevant genes (e.g., *APP* [40]) are excluded. Moreover, more famous and consolidated susceptibility genes (e.g., *APOE* [41], *BIN1* [42], *TNK1* [43] and so on.) seem to be filtered out, possibly due to the lower cohort of AD data in ADNI. In addition, the relationship between *RIN3* and AD can only be explained from the perspective of association analysis, and no further biological experiments are carried out to verify it.

(2) The occurrence and development of AD may not only be caused by SNPs, but also may be related to other mutations such as copy number variation. However, due to the data set analyzed and the computational framework designed in this work, the identified mutations or genes can only explain a part of AD/MCI samples. A more likely solution to this problem is that we will adopt the perspective of systems biology to supplement information from other levels of data (such as copy number variation, transcriptome, methylation, proteome and so on), and then integrate multi-level omics data to improve accuracy and sensitivity.

(3) XGBoost has certain advantages in classification on unbalanced data sets, but when the imbalance is very serious, its performance is still insufficient. In this work, the AD sensitivity of our method is low, because the sample size of AD is too small (only 46 AD samples). Further analysis revealed that this misclassification was mainly due to AD being wrongly identified as MCI, which indicates that the sample size is too small to accurately detect the pattern of AD. Therefore, in future research, we will study more effective sampling techniques or carry out larger-scale research to increase the sensitivity of the computational method.

Author contributions

JZ and XL carried out the design of the study and performed the statistical analysis.YPQ and XYL implemented the experiments and analyzed the results. YYP, SGL and ZRX participated in software coding and helped to draft the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Acknowledgment

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated com-pany Gen- entech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alz-heimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technolo-gies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Rev December 5, 2013 Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is

coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Funding

This work has been supported by the Jiangxi Provincial natural science fund (Nos. 20192ACB21004, 20204BCJL23035, 20212ABC03A32, 20202BAB212004 and 20212BAB202007), MOE (Ministry of Education in China) Project of Humanities and Social Sciences (No. 20YJAZH142) and Technological Research Project of Education Department in Jiangxi Province (GJJ190356 and GJJ210624).

Conflict of interest

The authors declare no conflict of interest.

References

[1] Sharma P, Srivastava P, Seth A, Tripathi PN, Banerjee AG, Shrivastava SK. Comprehensive review of mechanisms of pathogenesis involved in Alzheimer's disease and potential therapeutic strategies. Progress in Neurobiology. 2019; 174: 53–89.

[2] Tombaugh TN, McIntyre NJ. The mini-mental state examination: a comprehensive review. Journal of the American Geriatrics Society. 1992; 40: 922–935.

[3] Delacourte A, David JP, Sergeant N, Buee L, Wattez A, Vermersch P, *et al*. The biochemical pathway of neurofibrillary degeneration in aging and Alzheimer's disease. Neurology. 1999; 52: 1158–1158.

[4] Ren F, Yang C, Qiu Q, Zeng N, Cai C, Hou C, *et al*. Exploiting Discriminative Regions of Brain Slices Based on 2D CNNs for Alzheimer's Disease Classification. IEEE Access. 2019; 7: 181423–181433.

[5] Wang T, Xiao S, Liu Y, Lin Z, Su N, Li X, *et al*. The efficacy of plasma biomarkers in early diagnosis of Alzheimer's disease. International Journal of Geriatric Psychiatry. 2014; 29: 713–719.

[6] Xu L, Liang G, Liao C, Chen G, Chang C. K-Skip-n-Gram-RF: a Random Forest Based Method for Alzheimer's Disease Protein Identification. Frontiers in Genetics. 2019; 10: 33.

[7] Xu, L, Liang G, Liao CG, Chen GD, Chang CC. An Efficient Classifier for Alzheimer's Disease Genes Identification. Molecules. 2018; 23: 3140.

[8] Mormino EC, Sperling RA, Holmes AJ, Buckner RL, De Jager PL, Smoller JW, *et al*. Polygenic risk of Alzheimer disease is associated with early- and late-life processes. Neurology. 2016; 87: 481–488.

[9] Sun Q, Kong W, Mou X, Wang S. Transcriptional Regulation Analysis of Alzheimer's Disease Based on FastNCA Algorithm. Current Bioinformatics. 2019; 14: 771–782.

[10] Buniello A, MacArthur JA, Cerezo M, Harris LW, Hayhurst J, Malangone C, *et al*. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research. 2019; 47: D1005–D1012.

[11] Sun L, Liu G, Su L, Wang R. HS-MMGKG: a Fast Multi-objective Harmony Search Algorithm for Two-locus Model Detection in GWAS. Current Bioinformatics. 2019; 14: 749–761.

[12] Li P, Guo M, Wang C, Liu X, Zou Q. An overview of SNP interactions in genome-wide association studies. Briefings in Functional Genomics. 2015; 14: 143–155.

[13] Li X, Chen F, Xiao J, Chou S, Li X, He J. Genome-wide Analysis of the Distribution of Riboswitches and Function Analyses of the Corresponding Downstream Genes in Prokaryotes. Current Bioinformatics. 2019; 14: 53–61.

[14] Xu Z, Wu C, Pan W. Imaging-wide association study: Integrating imaging endophenotypes in GWAS. NeuroImage. 2017; 159: 159–169.

[15] Zeng X, Wang W, Deng G, Bing J, Zou Q. Prediction of Potential Disease-Associated MicroRNAs by Using Neural Networks. Molecular Therapy - Nucleic Acids. 2019; 16: 566–575.

[16] Wang L, Xuan Z, Zhou S, Kuang L, Pei T. A Novel Model for Predicting LncRNA-disease Associations based on the LncRNA-MiRNA-Disease Interactive Network. Current Bioinformatics. 2019; 14: 269–278.

[17] Kuang L, Zhao H, Wang L, Xuan Z, Pei T. A Novel Approach Based on Point Cut Set to Predict Associations of Diseases and LncRNAs. Current Bioinformatics. 2019; 14: 333–343.

[18] Zeng X, Liu L, Lü L, Zou Q. Prediction of potential disease-associated microRNAs using structural perturbation method. Bioinformatics. 2018; 34: 2425–2432.

[19] Sultana N, Sharma N, Sharma KP, Verma S. A Sequential Ensemble Model for Communicable Disease Forecasting. Current Bioinformatics. 2020; 15: 309–317.

[20] Pan W, Kim J, Zhang Y, Shen X, Wei P. A powerful and adaptive association test for rare variants. Genetics. 2014; 197: 1081–1095.

[21] Lambert JC, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, Combarros O, *et al*. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nature Genetics. 2009; 41: 1094–1099.

[22] Karch CM, Cruchaga C, Goate AM. Alzheimer's disease genetics: from the bench to the clinic. Neuron. 2014; 83: 11–26.

[23] Van Erp TG, Walton E, Hibar DP, Schmaal L, Jiang W, Glahn DC, *et al*. Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the Enhancing Neuro Imaging Genetics Through Meta Analysis (ENIGMA) Consortium. Biological Psychiatry. 2018; 84: 644–654.

[24] Shen L, Thompson PM, Potkin SG, Bertram L, Farrer LA, Foroud TM, *et al*. Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers. Brain Imaging and Behavior. 2014; 8: 183–207.

[25] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; 785–794.

[26] Lv Z, Wang D, Ding H, Zhong B, Xu L. Escherichia Coli DNA N-4-Methycytosine Site Prediction Accuracy Improved by Light Gradient Boosting Machine Feature Selection Technology. IEEE Access. 2020; 8: 14851–14859.

[27] Fu X, Cai L, Zeng X, Zou Q. StackCPPred: a stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency. Bioinformatics. 2020; 36: 3028–3034.

[28] Cai L, Ren X, Fu X, Peng L, Gao M, Zeng X. IEnhancer-XG: interpretable sequence-based enhancers and their strength predictor. Bioinformatics. 2020; 37: 1060–1067.

[29] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, *et al*. The Ensembl Variant Effect Predictor. Genome Biology. 2016; 17: 122.

[30] Zilberter Y, Gubkina O, Ivanov AI. A unique array of neuroprotective effects of pyruvate in neuropathology. Frontiers in Neuroscience. 2015; 9: 17.

[31] Oskouie AA, Yekta RF, Tavirani RM, Kashani MS, Goshadrou F. Lavandula angustifolia effects on rat models of alzheimer's disease through the investigation of serum metabolic features using NMR metabolomics. Avicenna Journal of Medical Biotechnology. 2018; 10: 83–92.

[32] Pereira CD, Martins F, Wiltfang J, da Cruz E Silva OAB, Rebelo S. ABC Transporters are Key Players in Alzheimer's Disease. Journal of Alzheimer's Disease. 2018; 61: 463–485.
[33] Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Research. 2001; 11: 863–874.
[34] Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, *et al*. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI) Alzheimer's & Dementia. 2005; 1: 55–66.
[35] Li X. A fast and exhaustive method for heterogeneity and epistasis analysis based on multi-objective optimization. Bioinformatics. 2017; 33: 2829–2836.
[36] Saini N, Saha S, Jangra A, Bhattacharyya P. Extractive single document summarization using multi-objective optimization: Exploring self-organized differential evolution, grey wolf optimizer and water cycle algorithm. Knowledge-Based Systems. 2019; 164: 45–67.
[37] Pathak GA, Silzer TK, Sun J, Zhou Z, Daniel AA, Johnson L, *et al*. Genome-Wide Methylation of Mild Cognitive Impairment in Mexican Americans Highlights Genes Involved in Synaptic Transport, Alzheimer's Disease-Precursor Phenotypes, and Metabolic Morbidities. Journal of Alzheimer's Disease. 2019; 72: 733–749.
[38] Shen R, Zhao X, He L, Ding Y, Xu W, Lin S, *et al*. Upregulation of RIN3 induces endosomal dysfunction in Alzheimer's disease. Translational Neurodegeneration. 2020; 9: 26.
[39] Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, *et al*. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Research. 2016; 44: W90–W97.
[40] Uruno A, Matsumaru D, Ryoke R, Saito R, Kadoguchi S, Saigusa D, *et al*. Nrf2 Suppresses Oxidative Stress and Inflammation in App Knock-In Alzheimer's Disease Model Mice. Molecular and Cellular Biology. 2020; 40: e00467–e00519.
[41] Serrano-Pozo A, Das S, Hyman BT. APOE and Alzheimer's disease: advances in genetics, pathophysiology, and therapeutic approaches. The Lancet Neurology. 2021; 20: 68–80.
[42] Gao P, Ye L, Cheng H, Li H. The Mechanistic Role of Bridging Integrator 1 (BIN1) in Alzheimer's Disease. Cellular and Molecular Neurobiology. 2021; 41: 1431–1440.
[43] Zeman T, Balcar VJ, Cahová K, Janoutová J, Janout V, Lochman J, *et al*. Polymorphism rs11867353 of Tyrosine Kinase Non-Receptor 1 (TNK1) Gene Is a Novel Genetic Marker for Alzheimer's Disease. Molecular Neurobiology. 2021; 58: 996–1005.