Electronic Research Archive

Research article

A modified FGL sparse canonical correlation analysis for the identification of Alzheimer's disease biomarkers

Shuaiqun Wang[1],*,[†], Huiqiu Chen[1],[†], Wei Kong[1], Xinqi Wu[1], Yafei Qian[1] and Kai Wei[2]

[1] College of Information Engineering, Shanghai Maritime University, 1550 Haigang Ave., Shanghai 201306, China
[2] Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China

* Correspondence: E-mail: wangsq@shmtu.edu.cn.

[†] These two authors contributed equally.

Abstract: Imaging genetics mainly finds the correlation between multiple datasets, such as imaging and genomics. Sparse canonical correlation analysis (SCCA) is regarded as a useful method that can find connections between specific genes, SNPs, and diseased brain regions. Fused pairwise group lasso-SCCA (FGL-SCCA) can discover the chain relationship of genetic variables within the same modality or the graphical relationship between images. However, it can only handle genetic and imaging data from a single modality. As Alzheimer's disease is a kind of complex and comprehensive disease, a single clinical indicator cannot accurately reflect the physiological process of the disease. It is urgent to find biomarkers that can reflect AD and more synthetically reflect the physiological function of disease development. In this study, we proposed a multimodal sparse canonical correlation analysis model FGL-JSCCAGNR combined FGL-SCCA and Joint SCCA (JSCCA) method which can process multimodal data. Based on the JSCCA algorithm, it imposes a GraphNet regularization penalty term and introduces a fusion pairwise group lasso (FGL), and a graph-guided pairwise group lasso (GGL) penalty term, the algorithm in this paper can combine data between different modalities, Finally, the Annual Depression Level Total Score (GDSCALE), Clinical Dementia Rating Scale (GLOBAL CDR), Functional Activity Questionnaire (FAQ) and Neuropsychiatric Symptom Questionnaire (NPI-

Q), these four clinical data are embedded in the model by linear regression as compensation information. Both simulation data and real data analysis show that when FGI-JSCCAGNR is applied to the imaging genetics study of Alzheimer's patients, the model presented here can detect more significant genetic variants and diseased brain regions. It provides a more robust theoretical basis for clinical researchers.

Keywords: imaging genetics; biomarkers; FGL-JSCCAGNR; regularization penalty term; clinical data;

Abbreviations: AD: Alzheimer's disease; SCCA: Sparse Canonical Correlation Analysis; CCA: Canonical Correlation Analysis; FGL: fused pairwise group lasso; GGL: graph-guided pairwise group lasso; CSs: cognitive scores; ADNI: Alzheimer's Disease Neuroimaging Initiative; sMRI: structural magnetic resonance imaging; ROIs: regions of interests; EMCI: Early Mild Cognitive Impairment; LMCI: Late Mild Cognitive Impairment; HC: Healthy Control; CCCs: canonical correlation coefficients; QTs: quantitative traits ; MRI: magnetic resonance imaging; SCD: subjective cognitive decline; GO: Gene Ontology; FDR: false discovery rate; KEGG: Kyoto Encyclopedia of Genes and Genomes; BAT: brown adipose tissue; Aβ: β-amyloid

1. Introduction

As a progressive neurodegenerative disease, Alzheimer's disease (AD) is characterized by an insidious onset [1]. AD occurs in presenile and elderly central nervous system degeneration. The symptoms are progressively aggravated, including cognitive decline and other neuropsychiatric symptoms. With the development of the disease, memory impairment gradually deteriorates, and other symptoms gradually appear. The disease cannot be cured, and comprehensive treatment can only alleviate the disease and delay its development.

According to a report released by the International Alzheimer's Association, an estimated 131.5 million people will suffer from AD by 2050.In future research, how to intervene and treat AD early, slow down the patient's disease's development, and find AD biomarkers have become an important topic in today's society. At the same time, the sparse canonical correlation analysis (SCCA) model plays a crucial role in diagnosing and predicting AD. These powerful methods can effectively leverage multiple data formats for neuroimaging and genomics [2,3].

Canonical Correlation Analysis (CCA) [4] can integrate multiple data types. However, current CCA-based fusion methods suffer from high-dimensional, multicollinearity in transforming imaging data into vectors. Mohammadi-Nejad AR et al. proposed a structured sparse CCA (ssCCA) to address the problems. It is finally shown that ssCCA outperforms existing standard regularization fusion methods [5]. Du et al. proposed GOSCAR-based SCCA, which encourages highly correlated features to have similar or equal canonical weights [6]. Du et al. proposed a new learning method and studied the multi-modal imaging genetic problem by constructing multiple SCCA tasks jointly, in which each associating SNPs with imaging QTs of one modality [7]. The FGL-SCCA model has been proposed, and it introduced two new penalty terms, which are fused pairwise group lasso (FGL) and graph-guided

pairwise group lasso (GGL) [8]. Nevertheless, it can only handle unimodal data, which limits the model. Based on FGL-SCCA, Qian et al. proposed the FGL-JSCCAR model [9] and extended it to a model that can handle multimodality while embedding MMSE into the model like linear regression. However, embedding only one clinical cognitive score has certain limitations on model performance. Hu et al. proposed sMCCA model for imaging genomics study and applied it to the analysis of SNPs, fMRI and methylation from schizophrenia studies. This research provided more ideas for future research on AD [10]. Wu et al. proposed the FGLGNSCCA model with Graphnet regularization to improve the efficiency and stability of SCCA and utilized it to deal with gene and sMRI data [11].

To address the appealing problem, this paper proposes a new FGl-JSCCAGNR model based on the JSCCA algorithm. First, imposing a GraphNet regularization penalty term, GraphNet verifies its stability and anti-jamming in the JCB-SCCA algorithm [12]. Second, this paper integrates multiple cognitive scores (CSs) as compensation information, and the training effect of the proposed model is more convincing than the uncertainty of the model trained with a single biological indicator. Gene data reflects the expression level of genes and SNP is a genetic marker formed by the variation of a single nucleotide in the genome. The same gene may cause inconsistent levels of gene expression due to SNP in different regions and its mutation. Therefore, this paper used two data of SNP and gene to explore the impact of AD-related brain regions from the micro and macro perspectives. The results of the model training in this paper show that our model can detect more significant genetic variants and diseased brain regions, which can also identify more specific biomarkers that provide a theoretical basis for the development of the disease, which has particular practical significance.

2. Methods

2.1. SCCA model

In the formulas in this article, bold lowercase letters represent vectors and bold uppercase letters describe matrices. Data set $X \in R^{n \times p}$, $Y \in R^{n \times q}$ in this article. X is the genotype data set as well as Y is the image data set. The CCA model is defined as follows:

$$\max_{u,v} \boldsymbol{u}^T \boldsymbol{X}^T \boldsymbol{Y} \boldsymbol{v} \tag{1}$$

$$s.t. \ \boldsymbol{u}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{u} = \boldsymbol{v}^T \boldsymbol{Y}^T \boldsymbol{Y} \boldsymbol{v} = 1,$$

Subsequently, Witten et al. proposed the SCCA model based on the above [13]. The definition is as follows:

$$\min_{u,v} -\boldsymbol{u}^T \boldsymbol{X}^T \boldsymbol{Y} \boldsymbol{v} + \lambda_u \|u\|_1^1 + \lambda_v \|v\|_1^1 \tag{2}$$

$$s.t. \ \|u\|_2^2 = \|v\|_2^2 = 1$$

2.2. FGL-SCCA model

Authors can cite a reference to the registration in the Materials and methods section. Du et al. imposed two new penalties, the FGL and the GGL. which have strong upper bounds on the grouping effects of both positively and negatively correlated variables:

$$\min_{u,v} - \boldsymbol{u}^T \boldsymbol{X}^T \boldsymbol{Y} \boldsymbol{v} + \Omega_{FGL}(\boldsymbol{u}) + \Omega_{GGL}(\boldsymbol{v}) \tag{3}$$

$$s.t. \ ||\boldsymbol{Xu}||^2 \le 1, ||\boldsymbol{Yv}||^2 \le 1,$$

Among them, the FGL and the GGL penalty terms are defined as:

$$\Omega_{FGL}(\boldsymbol{u}) = \lambda_1 \sum_{i=1}^{p-1} \omega_{i,i+1} \sqrt{u_i^2 + u_{i+1}^2}, \tag{4}$$

$$\Omega_{GGL}(\boldsymbol{v}) = \lambda_2 \sum_{(j,k) \in E} \omega_{j,k} \sqrt{v_j^2 + v_k^2} \tag{5}$$

where E is the edge guided by graph G, and $\omega_{j,k}$ represents the weight value of the edge.

2.3. FGL-JSCCAGNR model

In this paper, we propose a new structured sparse canonical correlation analysis method (FGL-JSCCAGNR). The model uses a connectivity-based penalty term, the GraphNet regularization term [14]. One of its advantages is that if the connectivity between the ith and jth nodes in the model is high, the GraphNet regularizer enforces that the corresponding elements of the canonical vector are similar, the formula is as follows:

$$P(\boldsymbol{u}) = \sum_{i,j} C_{\boldsymbol{u}(i,j)}(\boldsymbol{u}_i - \boldsymbol{u}_j)^2$$

$$P(\boldsymbol{v}) = \sum_{i,j} C_{\boldsymbol{v}(i,j)}(\boldsymbol{v}_i - \boldsymbol{v}_j)^2 \tag{6}$$

where $C_{\boldsymbol{u}(i,j)}$ and $C_{\boldsymbol{v}(i,j)}$ represent the connection of nodes i and j in $\boldsymbol{X}$ and $\boldsymbol{Y}$, and $\boldsymbol{u}_i$ and $\boldsymbol{u_j}$ represent the ith and jth elements of the standard vector, respectively. This article rewrites the penalty as:

$$P(\boldsymbol{u}) = \boldsymbol{u}^T L_u \boldsymbol{u}$$

$$P(\boldsymbol{v}) = \boldsymbol{v}^T L_v \boldsymbol{v} \tag{7}$$

where $\boldsymbol{L_u}$ and $\boldsymbol{L_v}$ are the Laplace matrices of the connection matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

At the same time, to achieve the purpose of combining data between different modalities, it should also be ensured that the graphical relationship or chain relationship in the same modal is not destroyed,

this paper introduces the FGL and the GGL penalty terms. Finally, we integrated multiple cognitive scores (CSs) into the model as compensation information.
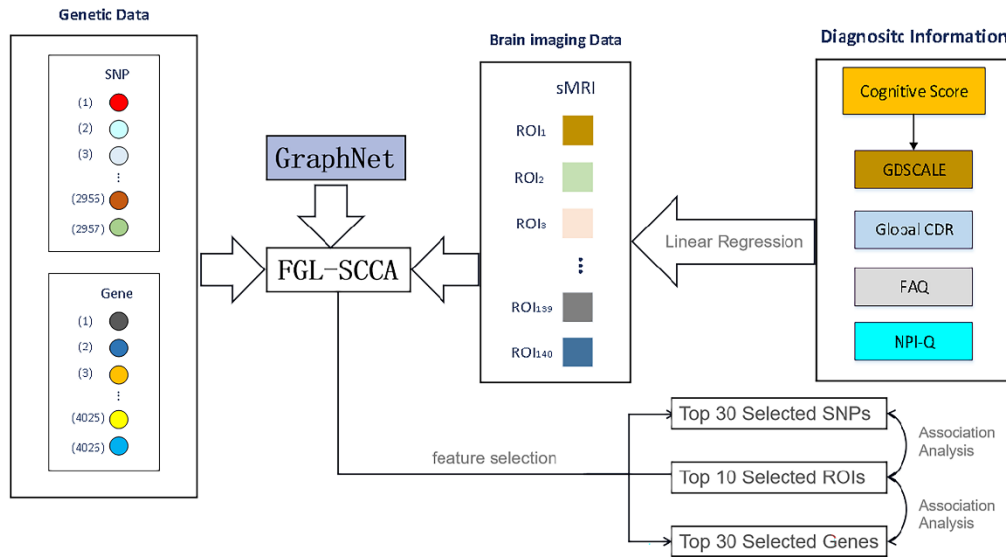

Figure 1. The schematic diagram of the proposed algorithm FGL-JSCCAGNR.

In the model FGL-JSCCAGNR constructed in this paper, $X_k \in R^{n \times p_k}$ and $Y \in R^{n \times q}$ represent the genetic variable matrix and the brain imaging variable matrix, u_k and v represent the feature weights of X_k and Y, respectively, $z_c \in R^{n \times 1} (c = 1, \dots, C)$ represents the clinical cognition score. The model algorithm is as follows:

$$min - \frac{1}{K} \sum_{k=1}^{K} u_k^T X_k^T Y v + \sum_{k=1}^{K} \Omega_{FGL}(u_k) + \Omega_{GGL}(v)$$

$$+ \frac{\gamma_1}{2} \sum_{k=1}^{K} (||X_k u_k||^2 - 1) + \frac{\gamma_2}{2} (||Yv||^2 - 1)$$

$$+ \frac{1}{2} \sum_{c=1}^{C} ||z_c - Yv||_2^2 + \frac{\lambda_3}{2} u^T L_u u + \frac{\lambda_4}{2} v^T L_v v \tag{8}$$

$$s.t. ||X_k u_k||^2 \le 1, ||Yv||^2 \le 1,$$

Among them, the definition of the expanded the FGL and the GGL penalty terms is:

$$\Omega_{FGL}(u_k) = \lambda_1 \sum_{i=1}^{p_k - 1} \omega_{i,i+1} \sqrt{u_{k,i}^2 + u_{k,i+1}^2}, \tag{9}$$

$$\Omega_{GGL}(\boldsymbol{v}) = \lambda_2 \sum_{m,n \in E} \omega_{m,n} \sqrt{v_m{}^2 + v_n{}^2} \,, \tag{10}$$

where K represents the total number of types of genetic variables, and k represents the kth modality. i, m and n represents the element's position in the vector. In the linear regression term, the vector $\boldsymbol{z}_c \in R'^{n \times 1} (c = 1, \dots, C)$ represents the clinical cognitive score, R' is an integer between 0 and 10. Figure 1 is the frame diagram of this model.

2.4. Surrogate objectives and efficient optimization algorithms

If the partial derivatives of $\boldsymbol{u}$ and $\boldsymbol{v}$ of Eq (8) are directly used by the Lagrangian function, the workload of direct calculation is very large. According to the conclusion drawn by Du in the literature, replace the multimodal data with function $\Omega_{FGL}^{APP}(\boldsymbol{u}_k)$, the form of $\Omega_{GGL}^{APP}(\boldsymbol{v})$ does not change. Let $||X_k \boldsymbol{u}_k||^2 = 1$ and $||Y\boldsymbol{v}||^2 = 1$, the Lagrangian function is as follows:

$$L(\boldsymbol{u}, \boldsymbol{v}) = -\frac{1}{K} \sum_{k=1}^{K} \boldsymbol{u}_k^T X_k^T Y \boldsymbol{v} + \sum_{k=1}^{K} \Omega_{FGL}^{APP}(\boldsymbol{u}_k) + \Omega_{GGL}^{APP}(\boldsymbol{v})$$

$$+ \frac{\gamma_1}{2} \sum_{k=1}^{K} (||X_k \boldsymbol{u}_k||^2 - 1) + \frac{\gamma_2}{2} (||Y\boldsymbol{v}||^2 - 1)$$

$$+ \frac{1}{2} \sum_{c=1}^{C} ||\boldsymbol{z}_c - Y\boldsymbol{v}||_2^2 + + \frac{\lambda_3}{2} \boldsymbol{u}^T L_u \boldsymbol{u} + \frac{\lambda_4}{2} \boldsymbol{v}^T L_v \boldsymbol{v} \tag{11}$$

γ_1 and γ_2 are positive tuning parameters, λ_1、λ_2、λ_3 and λ_4 are all tuning parameters. Since this Lagrangian function is continuous, both vectors $\boldsymbol{u}_k$ and $\boldsymbol{v}$ can be differentiated and are biconvex functions. It is only necessary to obtain partial derivatives for $\boldsymbol{u}_k$ and $\boldsymbol{v}$, then set $L(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{0}$, the extreme value can be obtained:

$$\boldsymbol{0} = -\frac{1}{K} X_k^T Y \boldsymbol{v} + \left(\lambda_1 \boldsymbol{D}_X \{k\} + \gamma_1 X_k^T X^T + \lambda_3 \boldsymbol{L}_u \right) \boldsymbol{u}_k \,, \tag{12}$$

$$\boldsymbol{0} = -\frac{1}{K} \sum_{k=1}^{K} Y^T X_k \boldsymbol{u}_k + [\lambda_2 \boldsymbol{D}_Y + (\gamma_2 + 1) Y^T Y + \lambda_4 \boldsymbol{L}_v] \boldsymbol{v} - \sum_{c=1}^{C} z_c Y^T \tag{13}$$

According to Eq (7), the alternating update equations of vectors and are obtained:

$$\boldsymbol{u}_k = \frac{\frac{1}{K} X_k^T Y \boldsymbol{v}}{\lambda_1 \boldsymbol{D}_X \{k\} + \gamma_1 X_k^T X_k + \lambda_3 \boldsymbol{L}_u} \tag{14}$$

$$\boldsymbol{v} = \frac{\frac{1}{K} \sum_{k=1}^{K} Y^T X_k \boldsymbol{u}_k + \sum_{c=1}^{C} z_c \, Y^T}{\lambda_2 \boldsymbol{D}_Y + (\gamma_2 + 1) Y^T Y + \lambda_4 \boldsymbol{L}_v}, \tag{15}$$

We show the calculation flow of the FGL-JSCCAGNR model in Table 1.

Table 1. Pseudo code for FGL-JSCCAGNR.

Algorithm 1: Algorithm for FGL-JSCCAGNR

Require: Normalized data $\mathbf{X}_k \in R^{n \times p_k}, \mathbf{Y} \in R^{n \times q}$,

$\mathbf{z}_c \in R'^{n \times 1}$, parameters λ_1, λ_2, λ_3, λ_4, γ_1, γ_2.

Ensure: Canonical vectors $\mathbf{u}_k, \mathbf{v}$.

1: Initialize $\mathbf{u}_k \in R^{p_k \times 1}, \mathbf{v} \in R^{q \times 1}$;

2: **while** not converged **do**

3: Update the diagonal matrix $\mathbf{D_X}\{k\}$, $P(\mathbf{u})$;

4: fix $\mathbf{V}$ and solve $\boldsymbol{u}_k = \dfrac{\frac{1}{K} X_k^T Y v}{\lambda_1 \boldsymbol{D}_X\{k\} + \gamma_1 X_k^T X_k + \lambda_3 \mathbf{L_u}}$;

5: scale $\boldsymbol{u}_k = \boldsymbol{u}_k./ sqrt(\boldsymbol{u}_k^T X_k^T X_k \boldsymbol{u}_k)$;

6: Update the diagonal matrix $\mathbf{D_Y}$, $P(v)$;

7: fix $\mathbf{u}_k$ and solve $\boldsymbol{v} = \dfrac{\frac{1}{K}\sum_{k=1}^{K} Y^T X_k u_k + \sum_{c=1}^{C} z_c Y^T}{\lambda_2 \boldsymbol{D}_Y + (\gamma_2 + 1) Y^T Y + \lambda_4 \mathbf{L}_v}$;

8: Scale $\mathbf{v} = \mathbf{v}./ sqrt(\mathbf{v}^T Y^T Y \mathbf{v})$;

9: **end while**

---

3. Results

3.1. Simulation data in FGL-JSCCAGNR

This paper uses simulation data to test the performance of the algorithm. First, three loading vectors are generated by simulation as ground truth, which represents the characteristics of genetic data and image data. Then set n samples, SNP data, and gene data are set to p = 600 feature dimension, and sMRI data is q = 90 dimensions. Since most of the four scales data values used in this algorithm are between 0 and 10, so in the setup, n integers in the interval 0 to 10 are arbitrarily selected and arranged into four-column vectors to replace the four clinical cognition data. To evaluate the anti-noise performance of the algorithm, we apply noise of different sizes to the generated data matrix, as shown in Figure 2, the canonical correlation coefficients (CCCs) of the four models all decrease with increasing applied noise. However, the figure shows that the FGL-JSCCAGNR curve decreases the slowest, and the noise impact is also minimal, it is confirmed that this model outperforms other models in terms of anti-noise performance.
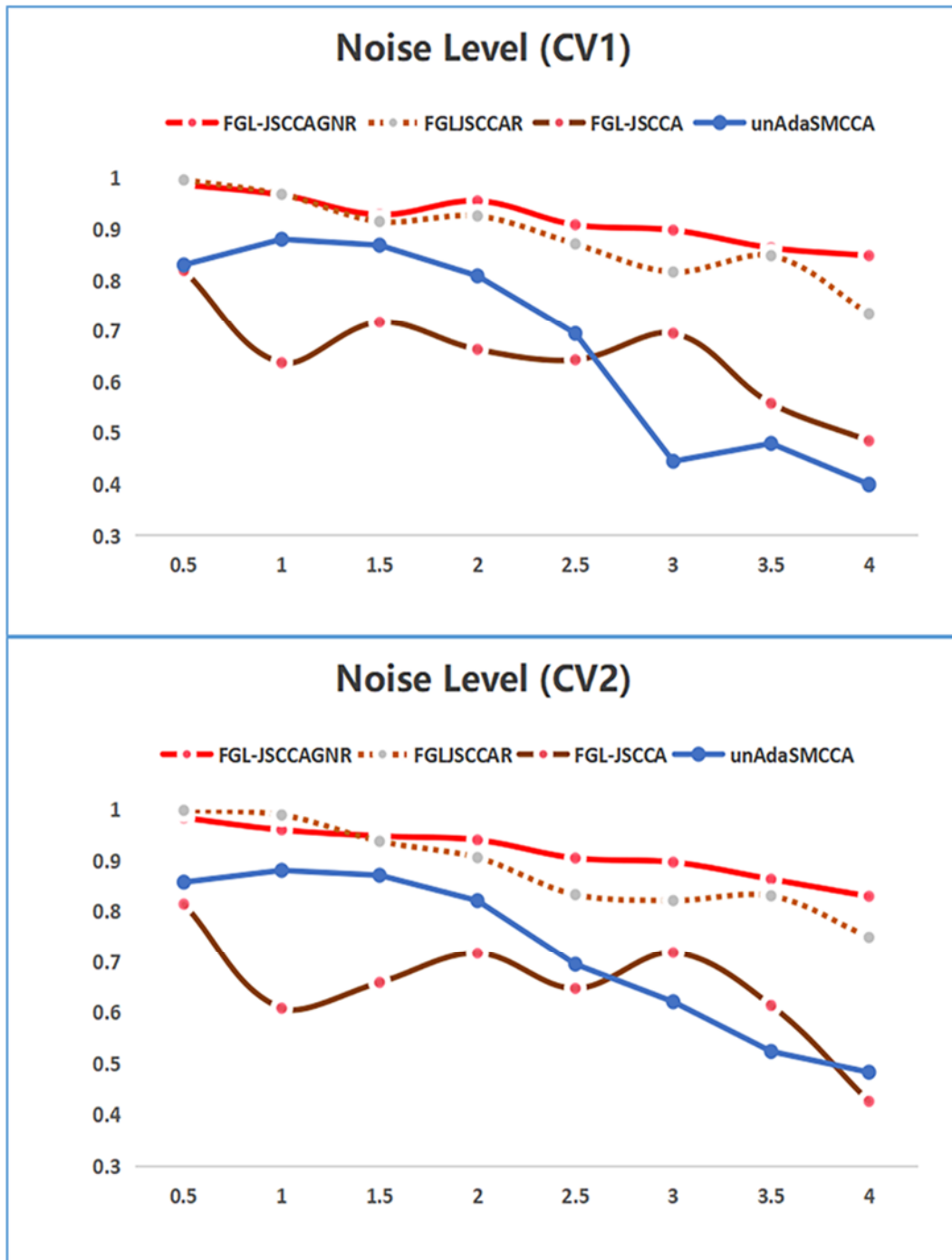
Figure 2. Comparison of CCCs under various noise levels for four models.

3.2. Subject data and preprocessing

The data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu) [15]. ADNI is primarily focused on connecting patients with data to determine the developmental stage of AD. The researchers recorded and preserved each patient's MRI, genetics, and cognitive test scores as predictors of disease. So far, the project has

compiled information on more than 1500 members, who are roughly 55–90 years old, and the data is available to all scientists.

We downloaded data in ADNI1 for 346 non-Hispanic white subjects, including SNP data, gene data, and sMRI data [16]. First, head motion-corrected structural magnetic resonance imaging (sMRI) was performed using DiffusionKit [17]. Next, sMRI segmentation was implemented using the CAT toolkit of the SPM software package [18], and the imaging phenotype features consisted of 140 regions of interest (ROI).

In this paper, PLINK [19] was used to preprocess SNP data, and the thresholds were set as HWE $p < 10^{-6}$, MAF < 0.05, and the difference calling rate of each SNP marker and each subject is set to be less than 90%, finally gender check and sibling pair identification are added, and 2957 SNP data are obtained.

At the same time, the four clinical cognitive score data of these samples were counted on ADNI, there are four clinical cognitive scores, namely, the GDSCALE, the GLOBAL CDR, the FAQ, and the NPI-Q, some scholars have confirmed that the scale is helpful for identifying biomarkers of AD. The subject characteristics statistics are shown in Table 2.

Table 2. Characteristics of the subjects.

Groups	AD	EMCI	LMCI	HC
Number	20	164	58	104
Gender (M/F)	13/7	90/74	30/28	54/50
Age (mean ± std)	77.81 ± 10.03	71.59 ± 7.52	73.06 ± 7.00	75.22 ± 5.82
GDSCALE (mean ± std)	1.90 ± 1.94	1.93 ± 1.85	1.98 ± 1.98	0.88 ± 1.42
Global CDR (mean ± std)	0.83 ± 0.37	0.54 ± 0.52	0.51 ± 0.11	0.03 ± 0.13
FAQ (mean ± std)	15.00 ± 6.51	2.05 ± 2.95	4.16 ± 4.82	0.19 ± 0.81
NPI-Q (mean ± std)	3.15 ± 2.92	2.57 ± 3.56	3.17 ± 3.28	0.68 ± 1.64

3.3. Experimental setup and parameter selection

In this study, we need to adjust a total of six parameters. Tuning all parameter values is not easy and is a time-consuming project. Where γ_v and γ_u mainly adjust the magnitude of V and U [20], so sets the value $\gamma_v = 1$, and $\gamma_u = 1$. In order to reduce the workload, we use a nested five-fold cross-validation method to match the values of the remaining parameters one by one from $(10^{-2}, 10^{-1}, 10^0, 10^1, 10^2)$.

This article uses the k-fold cross-validation method to train the dataset. After five cross-validations, we were able to obtain five correlation coefficients for gene-sMRI and SNP-sMRI and then average them:

$$CV = \frac{1}{5} \sum_{i=1}^{5} corr(\boldsymbol{Xt_i u O_i}, \boldsymbol{Yt_i v O_i}),$$

This article uses the CCCs of the highest mean test criterion that occurs as the choice. Taking the CCCs as an indicator, the purpose is to evaluate the model's performance. Meanwhile, CCCs can also be used as the Pearson correlation coefficient between Xu and Yv [21].
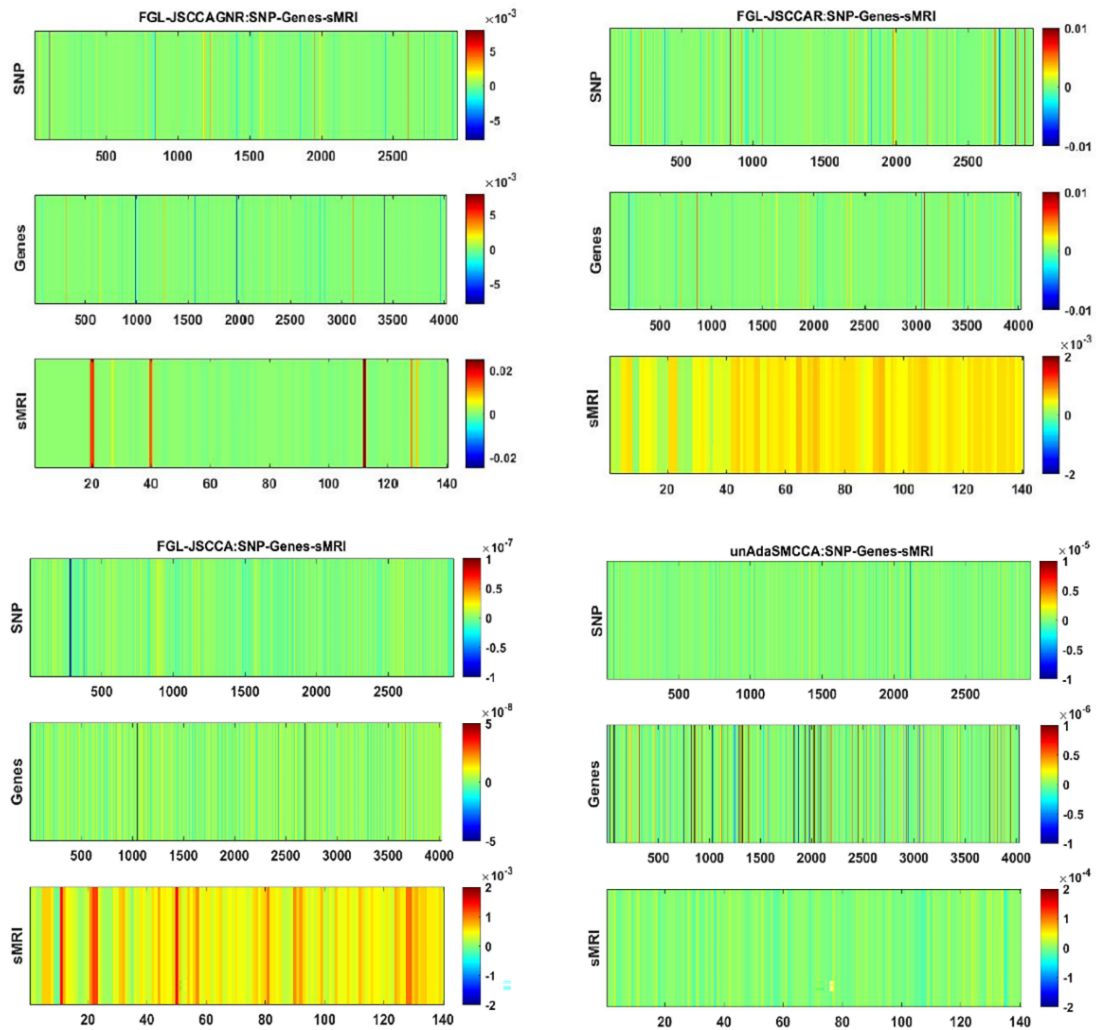
3.4. Experimental results

Figure 3. The heat maps obtained by different models.

Table 3. TOP10 Brain ROIs.

ROI	Weight
Left Occipital Pole	5.45E-02
Right Middle Temporal Gyrus	6.66E-04
Left Gyrus Rectus	5.85E-04
Left Inferior Temporal Gyrus	5.29E-04
Right Inferior Temporal Gyrus	5.20E-04
Right Middle Frontal Gyrus	5.09E-04
Left Planum Polare	5.06E-04
Left Cerebrum and Motor	4.88E-04
Right Anterior Insula	4.86E-04
Left Middle Frontal Gyrus	4.85E-04

Based on FGL-JSCCAR, this paper selected 346 sample data, including genetic data, imaging data, and four clinical cognitive score scales. Linear regression was used to select items related to the GDSCALE, the GLOBAL CDR, the FAQ, and the NPI-Q, these four clinical data strongly correlated with imaging quantitative traits (QTs). Training with 5-fold cross-validation produces a loading vector in a matrix. Finally, this paper will obtain three correlation matrices with SNP, gene, and sMRI data, and the matrix dimensions are 5 × 2957 dimensions, 5 × 4026 dimensions, and 5 × 140 dimensions, respectively. The above results, as shown in Figure 3.

Table 4. TOP30 SNP and Gene genetic feature weight.

SNP	Weight	Gene	Weight
rs12890143	4.21E-02	MYLK	2.67E-02
rs12911961	3.56E-02	SHC2	2.08E-02
rs12911832	3.69E-02	TMEM39A	1.91E-02
rs12912003	3.56E-02	NEK7	1.80E-02
rs6575275	3.47E-02	LOC100294033||FAM115A	1.73E-02
rs605928	3.17E-02	DUSP10	1.55E-02
rs56176704	2.94E-02	GPATCH11	1.55E-02
rs645275	2.73E-02	COLGALT2	1.31E-02
rs2041585553	2.61E-02	NDUFA1	1.27E-02
rs12435822	2.57E-02	RHBDD2	1.24E-02
rs10135622	2.54E-02	PLD3	1.21E-02
rs3795064	2.43E-02	CLVS1	1.16E-02
rs643010	2.23E-02	LOC100653057 || CES1	1.13E-02
rs769446	2.14E-02	SLC38A6	1.12E-02
rs12888973	2.13E-02	COMMD10	1.10E-02
rs10146142	1.99E-02	CMTM5	1.08E-02
rs181502870	1.98E-02	TCL6	1.07E-02
rs4807468	1.93E-02	TNFSF14	1.05E-02
rs77671856	1.92E-02	GEMIN5	1.05E-02
rs8177235	1.91E-02	CENPH	1.05E-02
rs663046	1.91E-02	PF4V1	1.03E-02
rs10141852	1.83E-02	NTNG2	9.99E-03
rs9671709	1.83E-02	ETV7	9.89E-03
rs4844384	1.82E-02	NCR3	9.63E-03
rs16979933	1.82E-02	KRTAP10-3	9.55E-03
rs4904975	1.73E-02	CARD16	9.36E-03
rs111278892	1.73E-02	WRN	9.32E-03
rs78877697	1.72E-02	LOC100128751	9.28E-03
rs34753032	1.70E-02	LOC202181 || SIMC1	8.73E-03
rs75627662	1.69E-02	CKB	8.73E-03

Figure 3 shows the standard weights for the real and genetic data. Table 3 lists the TOP10 brain regions obtained by this algorithm and the absolute value of the average weight obtained by 5 cross-validations. Due to the large dimension of genetic data, this paper selects the TOP30 SNPs and genes, and their average weights are shown in Table 4. To verify the effectiveness of Graphnet regularization in our proposed model, we compared it with FGLGNSCCA model. The CCCs of our model are higher than FGLGNSCCA mode which are listed in Tables 5 and 6. The larger the coefficient, the better the model effect. Table 5 lists the mean value and Standard Deviation (SD) of CCCs between SNP and sMRI with different models. Table 6 lists CCCs (mean ± SD) between genes and sMRI in different models. From these two tables, we found that the CCCs of our proposed model in this paper are better than the other three methods. From the perspective of biological significance, the model in this paper can identify a small number of SNPs related to AD from the huge and complex SNPs data, five of the top ten ROIs listed in Table 3 are brain risk regions associated with AD.

To highlight the regression effect of the four cognitive rating scales on the algorithm in this paper, this algorithm and the model without the scale were compared with other algorithms. This paper draws the Venn map of the top 10 brain regions obtained by these algorithms, as shown in Figure 4. We can conclude that when the algorithm in this paper regresses the scale, the top 10 brain regions obtained by the algorithm are more repetitive than those obtained by other algorithms, it is confirmed that the algorithm is better than the effect of no scale in the case of regression scale. The five models yielded a common brain region that is associated with AD. Compared with the other five models, the top10 brain regions identified by the models in this paper have the most identical brain regions. In addition to this, the model also found a brain region, the Right Anterior Insula, which those other algorithms had not picked up, which is likely to be a biomarker of AD.

Table 5. Canonical correlation coefficients of different models (SNP-sMRI).

model	CCCs (mean ± std）
FGL-JSCCAGNR	0.3711 ± 0.0604
FGL-JSCCAR	0.2905 ± 0.0121
unAdaSMCCA	0.3143 ± 0.0253
FGLGNSCCA	0.3412 ± 0.0264

Table 6. Canonical correlation coefficients of different models (gene-sMRI).

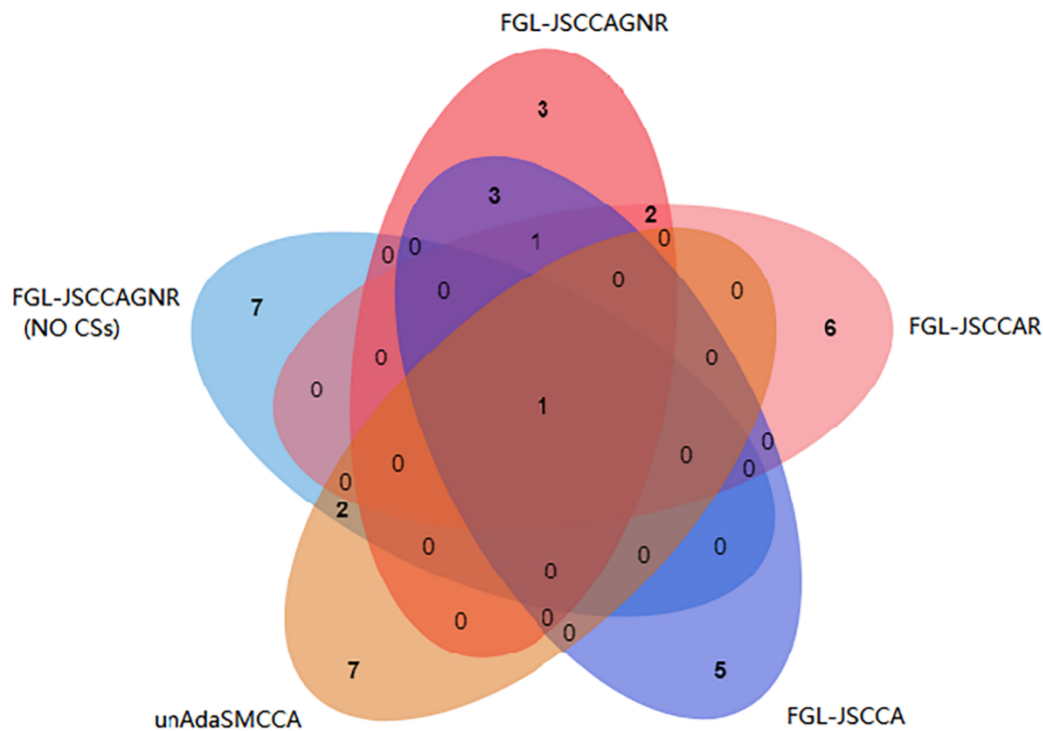| model	CCCs (mean ± std）
FGL-JSCCAGNR	0.3788 ± 0.0237
FGL-JSCCAR	0.3112 ± 0.0262
unAdaSMCCA	0.3328 ± 0.0149
FGLGNSCCA	0.3665 ± 0.0126

Figure 4. The venn diagram of the brain area.

4. Discussion

This paper used 346 samples, including genetic data and imaging data. The GraphNet regularization penalty term is imposed on the traditional model, and four clinical data of the GDSCALE, the GLOBAL CDR, the FAQ, and the NPI-Q are embedded into the algorithm as compensation information in a linear regression manner. This allows the model to predict diseased brain regions and biomarkers, allowing researchers to understand disease development more deeply.

4.1. Region of interest prediction

In ten brain regions, Left Occipital Pole, Right Middle Temporal Gyrus, and other brain regions have a great connection with AD. In a study that differentiated multimodal magnetic resonance imaging (MRI) from dementia subtypes in a traditional clinical setting, it was demonstrated that the index changes in the Left Occipital Pole could excellently separate healthy aging from neurodegenerative diseases [22]. Right Middle Temporal Gyrus confirmed in the article that RMTG hypometabolism may be a typical feature of subjective cognitive decline (SCD) and that large-scale hypometabolism in AD symptomatic patients may start from RMTG [23]. Blank F. et al. further investigated the neural basis of hallucinations in Alzheimer's disease (AD), leading to the conclusion

that the right anterior insula is hallucinogenic in cognitive neurodegenerative diseases core part [24]. Finally, related scholars also confirmed that the left inferior temporal gyrus has a certain influence on the diseases temporal lobe epilepsy [25] and schizophrenia [26], most likely as a result of complications of AD. The number of literature records in other brain regions is relatively small, and we need further confirmation, which can also bring researchers a new reference idea. Figure 5 is a visualization of the brain network of the above four abnormal brain regions and the shade of color indicates the weight of the identified brain regions.
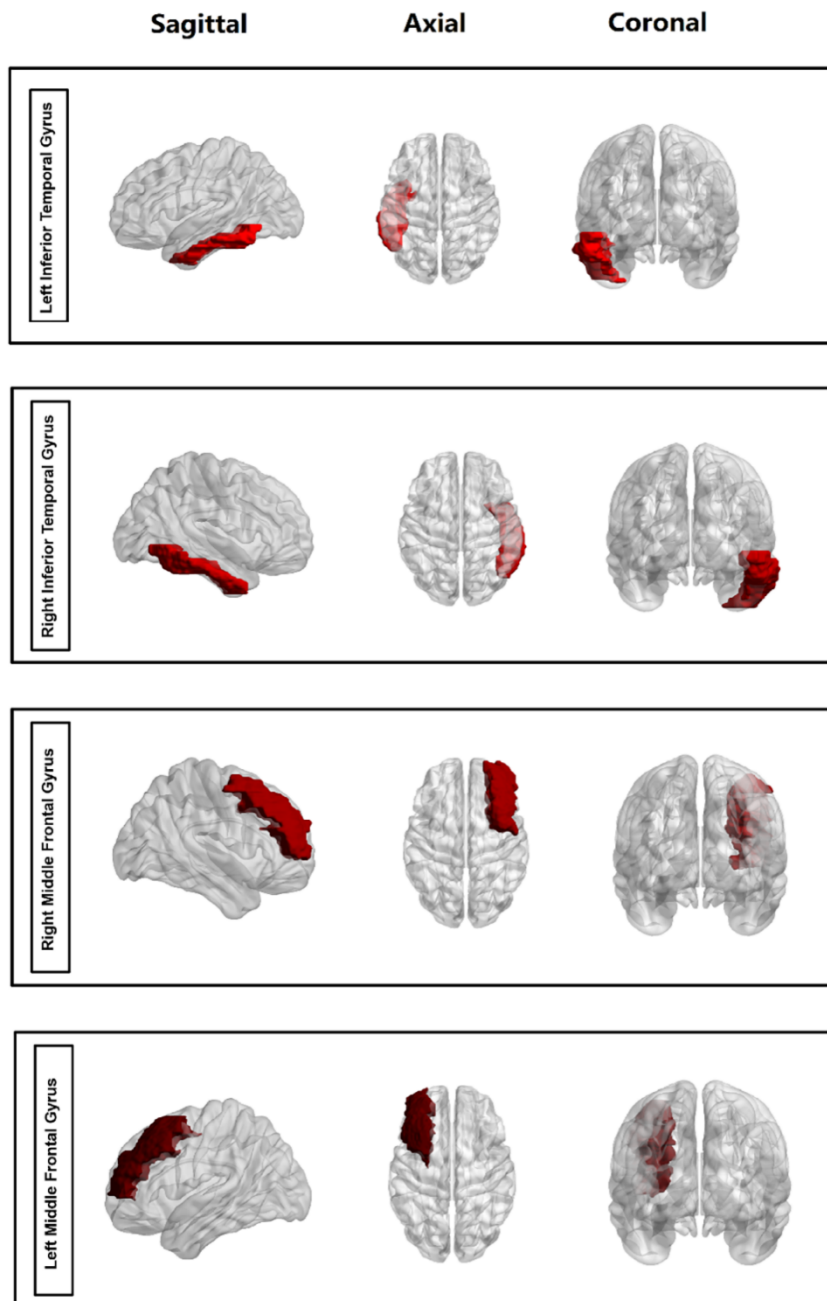

Figure 5. Visualization of the brain network of the four abnormal brain regions.

4.2. Over-Representation Analysis (ORA)

4.2.1. GO enrichment analysis

We first performed a Gene Ontology (GO) enrichment analysis of the top 600 genes using DAVID Bioinformatics Resources 6.8 (https://david.ncifcrf.gov/). We set a false discovery rate (FDR) < 0.05 as a cutoff to screening the data and the top 4 most representative terms after screening are shown in Table 7, where BP stands for biological process, and CC stands for a cellular component. Fifty-four genes were enriched in the four terms, and the detailed GO analysis chord diagram is shown in Figure 6.

Table 7. 4 sets of significant terms obtained by GO analysis.

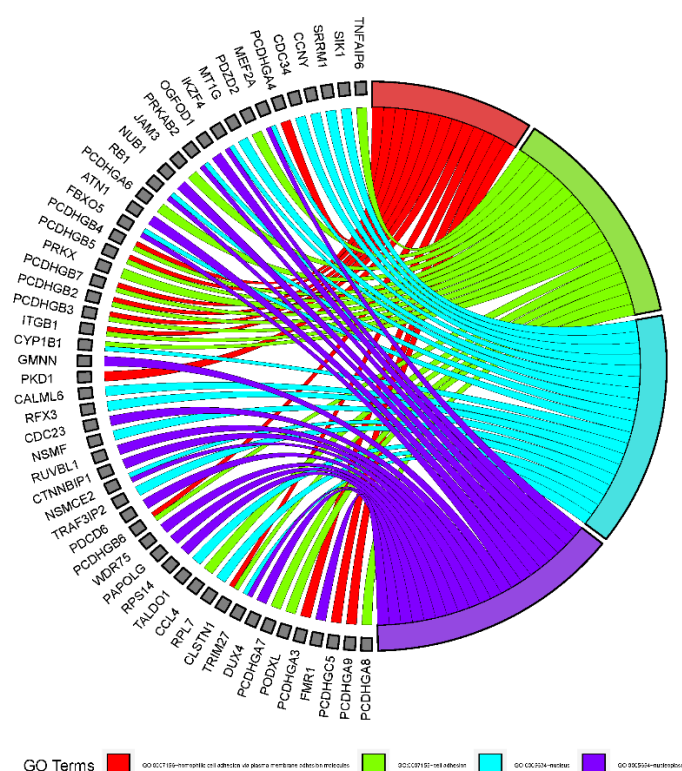| Category	ID	Term	FDR
BP	GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	4.03E-08
BP	GO:0007155	cell adhesion	4.31E-03
CC	GO:0005634	nucleus	1.27E-02
CC	GO:0005654	nucleoplasm	1.35E-02

Figure 6. Chord diagram obtained through GO analysis in the experiment. Among the first 600 genes entered, 55 genes were significantly enriched in 4 GO Term.

If a patient has AD, he is likely to be accompanied by other diseases. According to a description in one literature, AD affects the level and function of synaptic cell adhesion molecules, in particular, Aβ-dependent changes in adhesion to synapses affect synaptic function and integrity [27], Aβ is a common pathway for various reasons to induce AD, and it is also a key element in the pathogenesis of AD. The literature presents a systematic review of nuclear changes during AD development, including changes from the nuclear envelope to chromatin and epigenetic regulation. Finally, novel sequencing and gene perturbation techniques applied to address these challenges are introduced [28], which is likely to have important implications for future AD treatment. In addition, Läubli H et al. found that cell adhesion is critical in cancer-inducing mediators, including both immune evasion and metastatic spread [29], which is a fatal threat to patients and is not conducive to the treatment of AD.
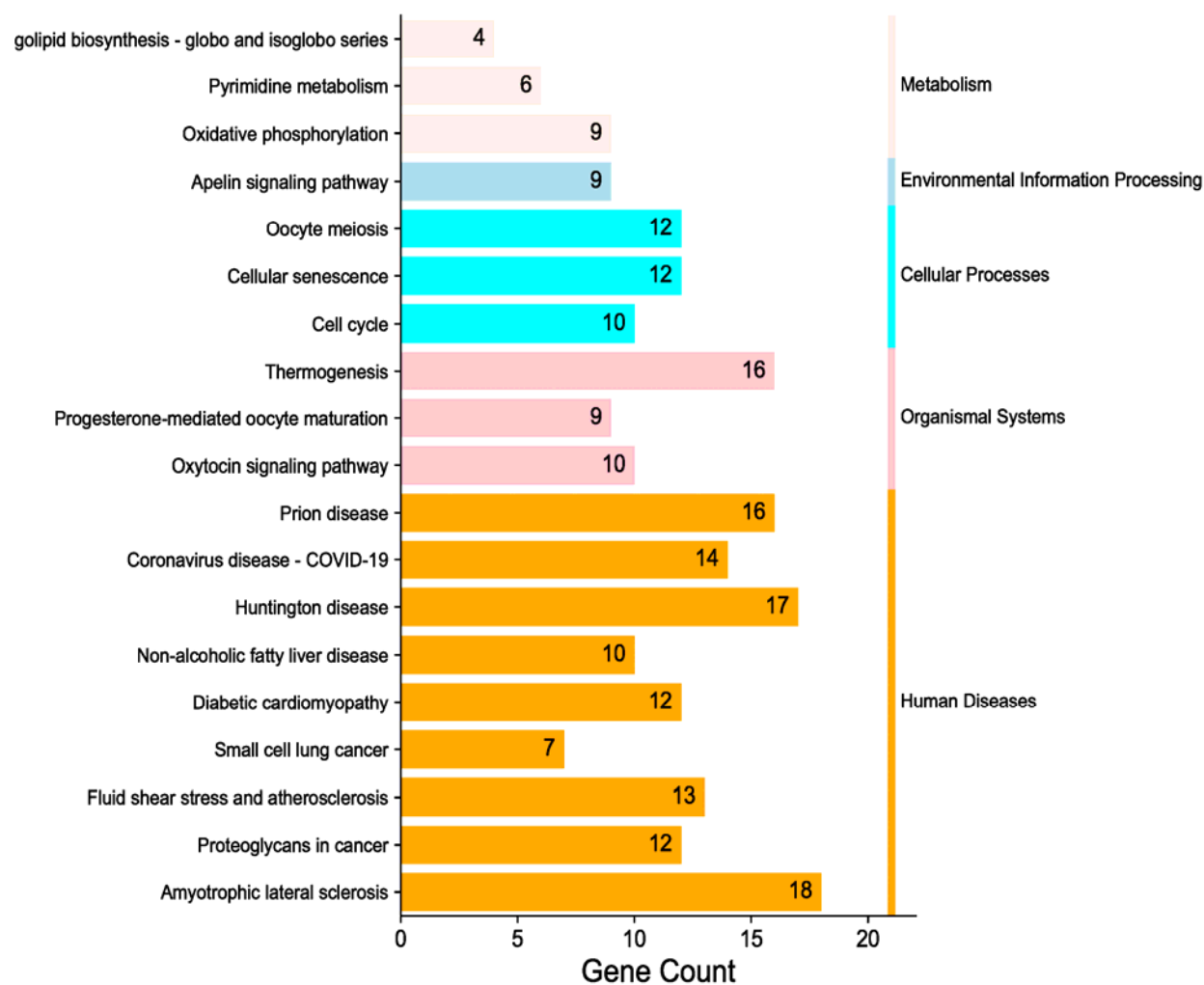
4.2.2. KEGG pathway analysis

Figure 7. KEGG enrichment analysis results of the first 600 genes.

We also performed the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis on the top 600 genes identified by the algorithm using the DAVID database. For all the information matched in the database, 261 IDs were enriched in the KEGG signaling pathway, accounting for 42%. The significant effect of the pathway was determined by the input gene enrichment effect, and we set a P-value < 0.05 as the critical value for screening signaling pathways to check the effectiveness of the pathway. As shown in Figure 7, 19 signaling pathways were obtained in this result.

In diseased cells, the activity of signaling pathways changes dramatically, affecting cell activity and making it difficult to treat the disease. The genes obtained by the algorithm in this paper have a great relationship with the related pathological processes of AD. Among them, the two signaling pathways, Thermogenesis and Huntington's disease, are better than other signaling pathways in terms of the number of enriched genes or their significance. Touristic M et al. suggest that human body thermoregulation, brown adipose tissue (BAT), and insulin-related metabolic deficits play a role in AD development and propose a mechanism by which correction of thermoregulatory dysfunction can slow AD progression and delay AD seizure [30]. Huntington's disease, like AD, is a neurodegenerative disease [31]. Perhaps relevant researchers can find the relationship between the two, which also provides a reference approach for the treatment of AD. Therefore, it also illustrates the superiority of our model, which can find AD-related signaling pathways and genes.

4.2.3. Refinement analysis

Among the SNP sites of TOP30 identified in this paper, there are 6 loci (rs56176704, rs12435822, rs10135622, rs181502870, rs77671856, rs10141852) that are located in SLC24A4, 5 loci (rs2041585553, rs605928, rs3795064, rs4807468, rs111278892) were located in the ABCA7 gene, and 3 loci (rs12911961, rs12911832, rs12912003) were located in the ADAM10 gene. Nettiksimmons J et al. found SLC24A4 associated with Alzheimer's disease (AD) in a large meta-analysis [32]. Fu Y et al. demonstrated that ABCA7 mediated the phagocytic clearance of amyloid-β in the brain and revealed a mechanism by which loss of ABCA7 function increases susceptibility to AD [33]. Yuan XZ et al. experimentally confirmed that ADAM10 could inhibit the production of β-amyloid (Aβ), and it can be used for tau pathology, synaptic function, and hippocampal neurogenesis through potential mechanisms. These contribute to the mitigation of AD pathological damage, both within the condition and outside the human body [34]. In addition, there are 7 loci (rs12890143, rs6575275, rs12888973, rs10146142, rs9671709, rs4904975, rs78877697) that are located in the RIN3 gene. In the article, among the 30 SNP loci obtained, the RIN3 gene appeared the most frequently. Xu Wei et al. confirmed that RIN3 (Ras and Rab interacting factor 3) was identified as one of the new risk factors for AD. As a guanine nucleotide exchange factor for Rab5, RIN3 may serve as its important activator in AD pathogenesis [35]. These scholars and the published literature all point out that the four genes listed in this paper have a great relationship with the pathological mechanism of AD, or that these genes directly or indirectly cause the pathogenesis of AD. ADAM10 and ABCA7 both affect AD by acting on amyloid-β. The model also identified TF, MEF2C, CR1, CASS4, and other gene loci, which have been confirmed in Kim JH's article, and they are all AD disease genes [36].

In addition, the pairwise correlation heatmaps of ROI-SNPs and ROI-genes are shown in Figures 8 and 9. As you can see, most of the ROI-SNP/gene pairs show a strong association. whether

it is a positive or negative correlation. To find significant correlations between genetic data and imaging data, we set P < 0.01 as the threshold and listed the top 10 ROI-SNP and ROI-gene pairs, as shown in Table 8. In addition, it was found that gene data showed a better correlation with ROI than SNP data. It can also provide ideas for later scholars to study biomarkers for identifying diseases from genetic data.
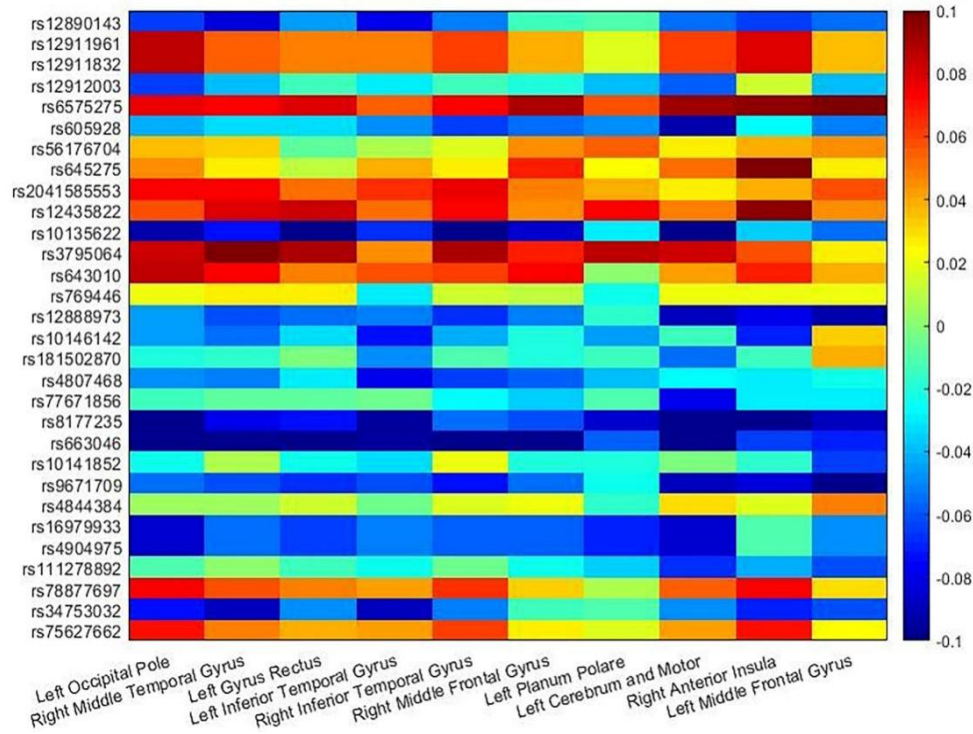

Figure 8. The paired correlation heat map of ROI-SNP.

Table 8. The top ten pairs of SNP-ROI and Gene-ROI with p < 0.05 in module 1, respectively.

| ROI-SNP	P Value	ROI-GENE	P Value
lOccPo - rs663046	1.13E-02	lOccPo - CKB	3.18E-03
rMidTemGy - rs663046	1.17E-02	rMidTemGy - PF4V1	2.23E-02
lRecGy - rs10135622	1.02E-02	lRecGy - NDUFA1	1.00E-02
lInfTemGy - rs8177235	1.95E-02	lInfTemGy - SHC2	2.73E-02
rInfTemGy - rs663046	1.20E-02	rInfTemGy - DUSP10	2.26E-02
rMidFroGy - rs4844384	1.92E-02	rMidFroGy - PF4V1	4.97E-02
lPla - rs12890143	1.09E-02	lPla - CKB	2.90E-02
lCbr+Mot - rs663046	1.28E-02	lCbr+Mot - LOC100294033 || FAM115A	8.64E-02
rAntIns - rs8177235	1.10E-02	rAntIns - CKB	1.11E-02
lMidFroGy - rs6575275	1.03E-02	lMidFroGy - SLC38A6	2.39E-02
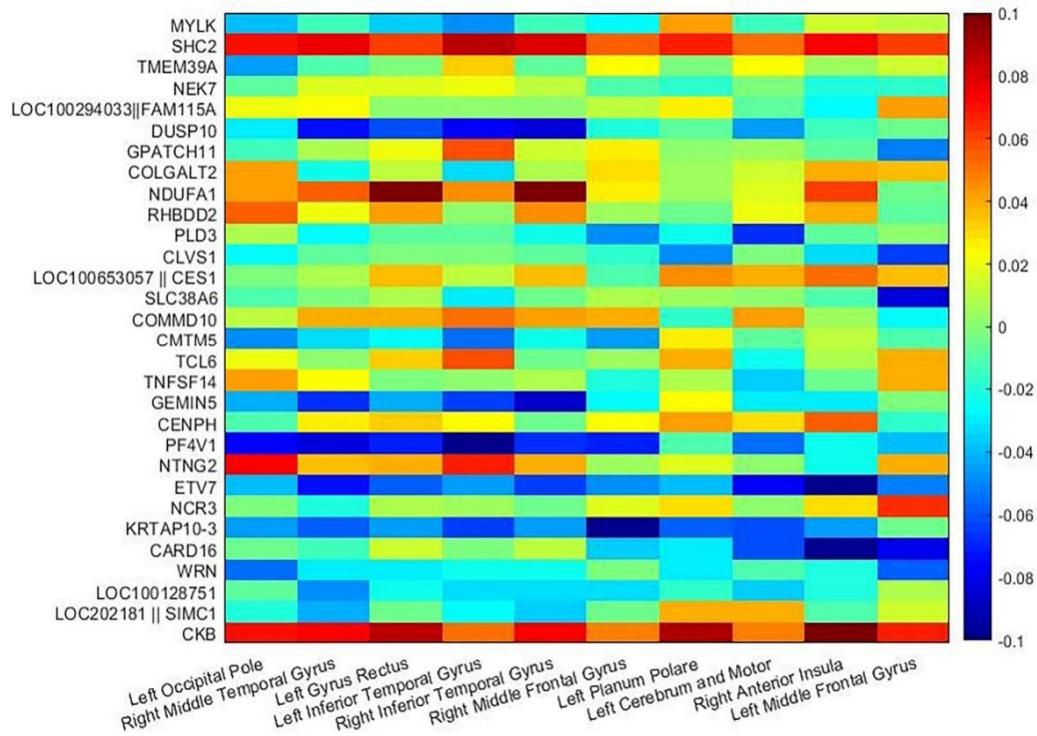
Figure 9. The paired correlation heat map of ROI-Gene.

5. Conclusions

This study utilized the efficient FGL-JSCCAGNR model to predict the diseased brain regions and biomarkers of AD. First, two penalty terms, FGL and GGL are imposed on JSCCA's multimodal data processing approach. Second, since changes in a single indicator in the dataset cannot measure the state of the disease, this paper collects and counts four cognitive score scales in a regression fashion to embed the model. Third, the GraphNet regularization term is added to the above model, and the final model is FGL-JSCCAGNR. Compared with other models, our model not only obtained higher correlation coefficients on the real dataset, identified more representative biomarkers, and when tested with simulated data, it also showed advantages that other models did not have. The pathogenesis of Alzheimer's disease is multifaceted. Using a single data or biomarker, derived algorithmic models are not reliable. It is hoped that the method proposed in this paper can provide some help for early AD patients to implement early intervention measures, and also provide researchers with new ideas to discover more reliable biomarkers. In future studies, we will try to integrate more imaging modalities (such as PET and fMRI) and more types of genetic data (such as DNA methylation data). It can also inject some new ideas into the study of AD, and it is hoped that the biological significance between brain regions and risk gene loci will be explored more accurately and comprehensively.

Acknowledgments

Conflict of interest

The authors declare that they have no competing interest.

References

1. N. Villain, B. Dubois, Alzheimer's disease including focal presentations, *Semin. Neurol.*, **39** (2019), 213–226. https://doi.org/10.1055/s-0039-1681041

2. M. Tanveer, B. Richhariya, R. Khan, A. Rashid, P. Khanna, M. Prasad, et al., Machine learning techniques for the diagnosis of Alzheimer's disease: A review, *ACM Trans.*, **16** (2020), 1–35. https://doi.org/10.1145/3344998

3. P. Khan, M. F. Kader, S. Islam, Machine learning and deep learning approaches for brain disease diagnosis: Principles and recent advances, *IEEE Access*, **9** (2021), 37622–37655. https://doi.org/10.1109/ACCESS.2021.3062484

4. E. Parkhomenko, D. Tritchler, J. Beyene, Sparse canonical correlation analysis with application to genomic data integration, *Stat. Appl. Genet. Mol. Biol.*, **8** (2009). https://doi.org/10.2202/1544-6115.1406

5. A. R. Mohammadinejad, G. A. HosseinZadeh, H. SoltanianZadeh, Structured and sparse canonical correlation analysis as a brain-wide multi-modal data fusion approach, *IEEE Trans. Med. Imaging*, **36** (2017), 1438–1448. https://doi.org/10.1109/TMI.2017.2681966

6. L. Du, H. Huang, J. Yan, S. Kim, S. Risacher, M. Inlow, et al., Structured sparse CCA for brain imaging genetics via graph OSCAR, *BMC Syst. Biol.*, **10** (2016), Supplement 3, 68. https://doi.org/10.1186/s12918-016-0312-1

7. L. Du, F. Liu, K. Liu, X. Yao, S. L. Risacher, J. Han, et al., Associating multi-modal brain Imaging phenotypes and genetic risk factors via a dirty multi-task learning method, *IEEE Trans. Med. Imaging*, **39** (2020), 3416–3428. https://doi.org/10.1109/TMI.2020.2995510

8. L. Du, J. Yan, S. Kim, S. L. Risacher, H. Huang, M. Inlow, et al., GN-SCCA: GraphNet based sparse canonical correlation analysis for brain imaging genetics, in *International Conference on Brain Informatics and Health*, **9250** (2015), 275–284. https://doi.org/10.1007/978-3-319-23344-4_27

9. S. Wang, Y. Qian, K. Wei, W. Kong, Identifying biomarkers of Alzheimer's disease via a novel structured sparse canonical correlation analysis approach, *J. Mol. Neurosci.*, **72** (2022), 323–335. https://doi.org/10.1007/s12031-021-01915-6

10. W. Hu, D. Lin, V. D. Calhoun, Y. P. Wang, Integration of SNPs-FMRI-methylation data with sparse multi-CCA for schizophrenia study, in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, **2016** (2016), 3310–3313. https://doi.org/10.1109/EMBC.2016.7591436

11. S. Wang, X. Wu, K. Wei, W. Kong, An improved fusion paired group lasso structured sparse canonical correlation analysis based on brain imaging genetics to identify biomarkers of Alzheimer's disease, *Front. Aging Neurosci.*, **13** (2022), 817520. https://doi.org/10.3389/fnagi.2021.817520

12. M. Kim, J. H. Won, J. Youn, H. Park, Joint-connectivity-based sparse canonical correlation analysis of imaging genetics for detecting biomarkers of Parkinson's disease, *IEEE Trans. Med. Imaging*, **39** (2020), 23–34. https://doi.org/10.1109/TMI.2019.2918839

13. L. Du, K. Liu, X. Yao, S. L. Risacher, J. Han, A. J. Saykin, et al., Detecting genetic associations with brain imaging phenotypes in Alzheimer's disease via a novel structured SCCA approach, *Med. Image Anal.*, **61** (2020), 101656. https://doi.org/10.1016/j.media.2020.101656

14. L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, J. E. Taylor, Interpretable whole-brain prediction analysis with GraphNet, *NeuroImage*, **72** (2013), 304–321. https://doi.org/10.1016/j.neuroimage.2012.12.062

15. M. W. Weiner, D. P. Veitch, Introduction to special issue: Overview of Alzheimer's disease neuroimaging initiative, *Alzheimer's Dementia*, **11** (2015), 730–733. https://doi.org/10.1016/j.jalz.2015.05.007

16. K. Wei, W. Kong, S. Wang, An improved multi-task sparse canonical correlation analysis of imaging genetics for detecting biomarkers of Alzheimer's disease, *IEEE Access*, **99** (2021), 30528–30538. https://doi.org/10.1109/ACCESS.2021.3059520

17. J. Gorski, F. Pfeuffer, K. Klamroth, Biconvex sets and optimization with biconvex functions: a survey and extensions, *Math. Methods Oper. Res.*, **66** (2007), 373–407. https://doi.org/10.1007/s00186-007-0161-1

18. A. J. Saykin, L. Shen, T. M. Foroud, S. G. Potkin, S. Swaminathan, S. Kim, et al., Alzheimer's disease neuroimaging initiative, Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans, *Alzheimer's Dementia*, **6** (2010), 265–273. https://doi.org/10.1016/j.jalz.2010.03.013

19. Y. Jung, J. Hu, A K-fold averaging cross-validation procedure, *J. Nonparam. Stat.*, **27** (2015), 167–179. https://doi.org/10.1080/10485252.2015.1010532

20. X. Chen, H. Liu, An effiffifficient optimization algorithm for structured sparse CCA, with applications to eQTL mapping, *Stat. Biosci.*, **4** (2011), 3–26. https://doi.org/10.1007/s12561-011-9048-z

21. X. Hao, C. Li, L. Du, X. Yao, J. Yan, S. L. Risacher, et al., Mining outcome-relevant brain imaging genetic associations via three-way sparse canonical correlation analysis in Alzheimer's disease, *Sci. Rep.*, **7** (2017), 44272. https://doi.org/10.1038/srep44272

22. T. Kuhn, S. Becerra, J. Duncan, N. Spivak, B. H. Dang, B. Habelhah, et al., Translating state-of-the-art brain magnetic resonance imaging (MRI) techniques into clinical practice: multimodal MRI differentiates dementia subtypes in a traditional clinical setting, *Quant. Imaging Med. Surg.*, **11** (2021), 4056–4073. https://doi.org/10.21037/qims-20-1355

23. Q. Y. Dong, T. R. Li, X. Y. Jiang, X. N. Wang, Y. Han, J. H. Jiang, Glucose metabolism in the right middle temporal gyrus could be a potential biomarker for subjective cognitive decline: a study of a Han population, *Alzheimer's Res. Ther.*, **13** (2021), 74. https://doi.org/10.1186/s13195-021-00811-w

24. F. Blanc, V. Noblet, N. Philippi, B. Cretin, J. Foucher, J. P. Armspach, et al., Right anterior insula: core region of hallucinations in cognitive neurodegenerative diseases, *PLoS One*, **9** (2014), e114774. https://doi.org/10.1371/journal.pone.0114774

25. K. Trimmel, A. L. van Graan, L. Caciagli, A. Haag, M. J. Koepp, P. J. Thompson, et al., Left temporal lobe language network connectivity in temporal lobe epilepsy, *Brain*, **141** (2018), 2406–2418. https://doi.org/10.1093/brain/awy164

26. T. Onitsuka, M. E. Shenton, D. F. Salisbury, C. C. Dickey, K. Kasai, S. K. Toner, et al., Middle and inferior temporal gyrus gray matter volume abnormalities in chronic schizophrenia: an MRI study, *Am. J. Psychiatry*, **161** (2004), 1603–1611. https://doi.org/10.1176/appi.ajp.161.9.1603

27. I. Leshchyns'ka, V. Sytnyk, Synaptic cell adhesion molecules in Alzheimer's disease, *Neural Plast.*, **2016** (2016), 6427537. https://doi.org/10.1155/2016/6427537

28. A. Iatrou, E. M. Clark, Y. Wang, Nuclear dynamics and stress responses in Alzheimer's disease, *Mol. Neurodegener.*, **16** (2021), 65. https://doi.org/10.1186/s13024-021-00489-6

29. H. Läubli, L. Borsig, Altered cell adhesion and glycosylation promote cancer immune suppression and metastasis, *Front. Immunol.*, **10** (2019), 2120. https://doi.org/10.3389/fimmu.2019.02120

30. M. Tournissac, M. Leclerc, J. Valentin-Escalera, M. Vandal, C. R. Bosoi, E. Planel, et al., Metabolic determinants of Alzheimer's disease: A focus on thermoregulation, *Ageing Res. Rev.*, **72** (2021), 101462. https://doi.org/10.1016/j.arr.2021.101462

31. W. Reith, Neurodegenerative diseases, *Radiologist*, **58** (2018), 241–258. https://doi.org/10.1007/s00117-018-0363-y

32. J. Nettiksimmons, G. Tranah, D. S. Evans, J. S. Yokoyama, K. Yaffe, Gene-based aggregate SNP associations between candidate AD genes and cognitive decline, *AGE*, **38** (2016), 41. https://doi.org/10.1007/s11357-016-9885-2

33. Y. Fu, J. H. Hsiao, G. Paxinos, G. M. Halliday, W. S. Kim, ABCA7 mediates phagocytic clearance of amyloid-β in the brain, *J. Alzheimer's Dis.*, **54** (2016), 569–84. https://doi.org/10.3233/JAD-160456

34. X. Z. Yuan, S. Sun, C. C. Tan, J. T. Yu, L. Tan, The role of ADAM10 in Alzheimer's disease, *J. Alzheimer's Dis.*, **58** (2017), 303–322. https://doi.org/10.3233/JAD-170061

35. W. Xu, F. Fang, J. Ding, C. Wu, Dysregulation of Rab5-mediated endocytic pathways in Alzheimer's disease, *Traffic*, **19** (2018), 253–262. https://doi.org/10.1111/tra.12547

36. J. H. Kim, Genetics of Alzheimer's disease, *Dementia Neurocognitive Disord.*, **17** (2018), 131–136. https://doi.org/10.12779/dnd.2018.17.4.131