IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 28, 2020, accepted February 10, 2021, date of publication February 15, 2021, date of current version February 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3059520

An Improved Multi-Task Sparse Canonical Correlation Analysis of Imaging Genetics for Detecting Biomarkers of Alzheimer's Disease

KAI WEI, WEI KONG, AND SHUAIQUN WANG
College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China
Corresponding author: Wei Kong (weikong@shmtu.edu.cn)

This work was supported in part by the Natural Science Foundation of Shanghai under Grant 18ZR1417200, and in part by the National Natural Science Foundation of China under Grant 61803257.

ABSTRACT Imaging genetics research based on Sparse Canonical Correlation Analysis (SCCA) helps to discover the correlation between pathological features reflected by neuroimaging and genotypic variation. Multi-Task SCCA (MTSCCA) method is to identify bi-multivariate associations between SNPs and multi-modal imaging QTs. However, the MTSCCA method is unsupervised and cannot identify diagnosis-guided genotype-phenotype associations. In order to improve the performance and interpretability of MTSCCA, we propose an improved MTSCCA algorithm, which is a supervised sparse bivariate learning model fused with a linear regression model, in which the regression part plays a guiding role in imaging QT selection. To jointly understand the relationship between genotypes and phenotypes of multiple tasks, in this study, gene expression data and single nucleotide polymorphisms (SNPs) as genetic data are considered in the algorithm. The focus of each task is to determine the genotype-phenotype pattern guided by the diagnostic team to discover the association with SNP/gene and brain region changes. Besides, the Laplacian matrices of three kinds of data are added as prior knowledge to the algorithm penalty item so that the algorithm can analyze the correlation between different features. Compared with other SCCA methods, our algorithm has improved noise resistance and stability, and found some diagnostic-specific SNP/gene-ROI specific to the two diagnostic groups of MCI and AD. Significance: This method provides a way to further study the association of multi-modal biological data and identify the complex association patterns of diseases.

INDEX TERMS Alzheimer's disease (AD), imaging genetics, sparse canonical correlation analysis (SCCA).

I. INTRODUCTION

In recent years, imaging genetics has become an important research topic because of its ability to explore the effects of genes on the structure and function of the brain and reveal the pathogenesis of some brain diseases. Image genetics has made significant progress in studying the pathogenesis of Alzheimer's disease (AD) and mining biomarkers.

To make up for the low accuracy problem of early imaging genetics association relationship analysis based on univariate paired statistical analysis methods, in recent years, image genetics based on statistical learning has enabled the association analysis of multivariate genetic data with quantitative traits in multivariate imaging. How to effectively use the high-dimensional genetic sites and sparse features of images has become a research hotspot in biological data association analysis.

Canonical Correlation Analysis (CCA) [1] is a standard multivariate method that integrates two or more data types. It can maximize the linear combination of the most remarkable correlation among different types of variables and then obtain the interrelated components of the two sets of data. Therefore, when using traditional CCA methods, severe overfitting may occur. In response to this problem, the sparse-constrained CCA method [2], [3] was introduced, which can identify bivariate associations between multiple SNPs

The associate editor coordinating the review of this manuscript and approving it for publication was Navanietha Krishnaraj Krishnaraj Rathinam.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

and multiple imaging QTs. Many scholars made improvements based on SCCA. Considering the different group-level structures of genetic data and imaging traits, some scholars improved the LASSO penalty that combines LASSO and graph/network guidance in structured sparse learning. For example, to balance the weighted similarity between grouping features or associated features, the KG-SCCA method [4] is proposed. It can model two types of prior knowledge, one as a group structure and the other as a network structure. DU *et al.* introduced a new graph-guided format to solve the traditional sparse feature selection problem in the case of incomplete prior knowledge [5]. Hao *et al.* also proposed a time-constrained longitudinal image gene association to study and explore the dynamic relationship between multiple genetic variables and multiple image markers [6]. Multi-task SCCA (multi-task-SCCA, MTSCCA) is recently proposed to study the genetic problems of multimodal imaging by jointly constructing multiple SCCA tasks in the literature [7]. The newly proposed MT-SCCALR [8] is superior to its sophisticated modeling strategy, which enables it to identify the characteristics of the diagnostic group and is of great significance for clinical research. However, due to its algorithm's time-consuming and program limitations, it is not suitable for large-scale association analysis, also. Fang *et al.* modified based on mSCCA and then proposed JSCCA to study the association of each imaging genetic group. However, its disadvantage is that it does not add a priori matrix to consider the different internal characteristics of the two modal data. When it performs SCCA in a single diagnostic group, many undesirable associations may dominate the associations of interest [9].

To solve these problems, this paper proposes an improved MTSCCA algorithm. First, add gene expression data parallel to SNP as genetic data to explore the influence of multimodal genetic data on brain structure. Second, the GraphNet regularizer [10] is added to the penalty term. The GraphNet regularizer is a modified version of elastic net regularization, which can effectively integrate physiological constraints such as connectivity. Its stability and anti-interference have been proved in the JCB-SCCA algorithm [11]. Using it as a priori knowledge into the algorithm makes the results of the algorithm more biologically explanatory. Third, we integrate the linear regression model into the MTSCCA algorithm, so that the diagnosis information can be added, and then the diagnosis-guided SNP/gene-ROI belonging to diagnosis groups are obtained. The results show that the proposed algorithm has better anti-noise performance in simulated data sets, and higher canonical correlation coefficients and classification accuracy in most diagnostic groups of real data sets.

II. METHOD

A. SPARSE CANONICAL CORRELATION ANALYSIS (SCCA)

CCA is a method to determine the association between two data sets. Given a data set $\boldsymbol{X} \in R^{n\times p}$, $\boldsymbol{Y} \in R^{n\times q}$, where $\boldsymbol{X}$ has p features, and $\boldsymbol{Y}$ has q features, a total of n samples. $\boldsymbol{X}$ represents the p feature of the imaging data, and $\boldsymbol{Y}$ represents the q feature of the SNP data and q feature of the Gene expression data. Using the CCA method can achieve high-dimensional data $\boldsymbol{X}$, $\boldsymbol{Y}$ dimensionality reduction, in this process to find a linear combination of variables in $\boldsymbol{X}$ and $\boldsymbol{Y}$ to maximize correlation.

$$\begin{aligned} &\max_{u,v} u^T X^T Yv \\ &\text{s.t. } u^T X^T Xu = v^T Y^T Yv = 1 \end{aligned} \tag{1}$$

Among them, we assume that the columns of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are standardized. The mean and unit variance is zero. u and v are the corresponding norm vectors.

The SCCA model was proposed by Parkhomenko *et al.* [2], Du *et al.* [3] based on CCA. The innovation lies in the proposed l_1 penalty for controlling the sparsity of the model, which is defined as follows:

$$\begin{aligned} &\min_{\mathrm{u,v}} -u^T X^T Yv + \lambda_{\mathrm{u}} \|u\|_1^1 + \lambda_{\mathrm{v}} \|v\|_1^1 \\ &\text{s.t. } \|u\|_2^2 = \|v\|_2^2 = 1 \end{aligned} \tag{2}$$

where u and v are corresponding norm vectors.

B. AN IMPROVED MTSCCA

The MTSCCA model was proposed by Du *et al.* We innovatively added gene expression data in parallel with SNP to genetic data. c represents the number of diagnostic groups. Specifically, we use $\boldsymbol{X_{cj}} \in R^{n\times p}$ to represent the phenotype data with p imaging, and $\boldsymbol{Y_{cj}} \in R^{n\times q}$ $(j = 1, 2)$ to represent the genetic data with q SNPs, q gene expression data from c diagnostic groups. Let $\mathbf{U} \in R^{p\times c}$ be the canonical weight matrix associated with $\boldsymbol{X_{cj}}$ and $\boldsymbol{V_j} \in R^{q\times c}$ be that associated with $\boldsymbol{Y_{cj}}$, the MTSCCA model is as follows.

$$\begin{aligned} &\min_{u,v_j} \sum_{c=1}^{n} \sum_{j=1}^{2} -u_c^T X_c^T Y_{cj} v_{cj} \\ &s.t.\ \|X_c u_c\|_2^2 = 1,\ \left\|Y_{cj} v_{cj}\right\|_2^2 = 1, \\ &\quad \Omega(U) \le b_1, \quad \Omega\left(V_j\right) \le b_2 \end{aligned} \tag{3}$$

To encourage the similarity of related elements of the norm vector, the connectivity-based penalty term [11] is introduced in our algorithm. Specifically, if the connectivity between the i-th node and the j-th node (that is, the brain region or SNP site) is high, it will force the corresponding elements of the norm vector to be similar. Therefore, we add the information of the brain connectivity and the weighted SNP correlation network to capture the genetic network structure [12] as a prior matrix to add to the algorithm, aiming to improve the biological significance of the extracted features.

$$\begin{aligned} P(u) &= \sum_{p,q} L_{u(p,q)}(u_p - u_q) \\ P(v) &= \sum_{p,q} L_{v(p,q)}(v_p - v_q) \end{aligned} \tag{4}$$

Here, L_u and L_v represent the Laplacian matrix of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively.

We regard the Laplacian matrix of $\boldsymbol{X}$ and $\boldsymbol{Y}$ as a new fusion penalty of MTSCCA algorithm, and the specific formula is as follows:

$$\begin{aligned} P(u) &= u^T L_u u \\ P(v) &= v^T L_v v \end{aligned} \tag{5}$$

L_u and L_v represent the matrix of $\boldsymbol{X}$ and $\boldsymbol{Y}_j$, respectively. The Laplacian matrices are defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{C}$, where $\boldsymbol{D}$ is the degree matrix of connectivity matrix $\boldsymbol{P}$.

Then, we use linear regression to identify relevant ROI. Since the association between QTs and genetic data will eventually encourage the identified QTs to be distinguished from each other, we only include the regression terms between QTs and category labels.

$$\min_{u} \sum_{c=1}^{n} \frac{1}{c} \|z_c - X_c u_c\|$$

z_c is the corresponding class label of the lth subject for the cth task.

C. THE EFFICIENT OPTIMIZATION ALGORITHM

Now we can write the proposed method with penalties explicitly exhibited.

$$\begin{aligned} &\min_{u,v_j} \sum_{c=1}^{n} \sum_{j=1}^{2} -u_c^T X_c^T Y_{cj} v_{cj} + \sum_{c=1}^{n} \frac{1}{n} \|z_c - X_c u_c\| \\ &s.t.\ \|X_c u_c\|_2^2 = 1, \left\|Y_{cj} v_{cj}\right\|_2^2 = 1, P(u_c) \le d_1, \\ &P\left(v_{cj}\right) \le d_2, \|u_c\|_1 = d_3, \left\|v_{cj}\right\|_1 = d_4 \end{aligned} \tag{6}$$

We modify the loss function to

$$\begin{aligned} &\min_{u,v_j} \sum_{c=1}^{n} \sum_{j=1}^{2} \left\|X_c u_c - Y_{cj} v_{cj}\right\|_2^2 + \sum_{c=1}^{n} \frac{1}{n} \|z_c - X_c u_c\| \\ &s.t.\ \|X_c u_c\|_2^2 = 1, \left\|Y_{cj} v_{cj}\right\|_2^2 = 1, P(u_c) \le d_1, \\ &P\left(v_{cj}\right) \le d_2, \|u_c\|_1 = d_3, \left\|v_{cj}\right\|_1 = d_4 \end{aligned} \tag{7}$$

which is equivalent to the original one, $\|X_c u_c\|_2^2 = 1$, $\left\|Y_{cj} v_{cj}\right\|_2^2 = 1$. Then we write its Lagrangian.

$$\begin{aligned} &L(u, v) = \sum\nolimits_{c=1}^{n} \frac{1}{n} \|z_c - X_c u_c\| + \sum\nolimits_{c=1}^{n} \sum\nolimits_{j=1}^{2} \\ &\left\|X_c u_c - Y_{cj} v_{cj}\right\|_2^2 + \gamma_1 \left(\|X_c u_c\|_2^2 - 1\right) \\ &+\gamma_2 \left(\left\|Y_{cj} v_{cj}\right\|_2^2 - 1\right) + \beta_1 \left(\|u_c\|_1 - d_3\right) \\ &+\beta_2 \left(\left\|v_{cj}\right\|_1 - d_4\right) + \lambda_1 \left(P(u_c) - d_1\right) \\ &+\lambda_2 \left(P\left(v_{cj}\right) - d_2\right) \end{aligned} \tag{8}$$

where β_1, β_2, λ_1, λ_2, γ_1 and γ_2 are tuning parameters, and β_1, β_2, λ_1 and λ_2 are positive values which control the model sparsity. By dropping the constants, we further have

$$\begin{aligned} &L(u, v) = \sum\nolimits_{c=1}^{n} \frac{1}{n} \|z_c - X_c u_c\| + \sum\nolimits_{c=1}^{n} \sum\nolimits_{j=1}^{2} \\ &\left\|X_c u_c - Y_{cj} v_{cj}\right\|_2^2 + \gamma_1 \|X_c u_c\|_2^2 \\ &+\gamma_2 \left\|Y_{cj} v_{cj}\right\|_2^2 + \beta_1 \|u_c\|_1 + \beta_2 \left\|v_{cj}\right\|_1 \\ &+\lambda_1 P(u_c) + \lambda_2 P\left(v_{cj}\right) \end{aligned} \tag{9}$$

from the point of view of optimization.

In order to minimize the objective function (9) to obtain the optimal u and v algorithm, we use the alternating convex search method [13]. First, we initialize u and v, then, when u is fixed, we modify v, and vice versa. And repeat the above process until convergence. We apply [2] to solve the minimization problem.

$$\begin{aligned} &\min_{v} \sum_{c=1}^{n} \sum_{j=1}^{2} \left\|X_c u_c - Y_{cj} v_{cj}\right\|_2^2 + \beta_2 \left\|v_{cj}\right\|_1 \\ &+\lambda_2 \boldsymbol{v}_{cj}^T L_{vcj} \boldsymbol{v}_{cj} + \gamma_2 \left\|Y_{cj} v_{cj}\right\|_2^2 \\ &s.t.\ \left\|Y_{cj} v_{cj}\right\|_2^2 = 1, \left\|v_{cj}\right\|_1 = d_4 \end{aligned} \tag{10}$$

We first get a value of v, which is minimized when u is constant. Since the Laplacian matrix is a positive definite matrix, the penalties based on connectivity are convex, that is, equation (10) is convex and can be optimized using coordinates based on soft thresholds, so the coordinate solution of (10) is defined as

$$v_{cj} = H(\sum_{c=1}^{n} \sum_{j=1}^{2} [Y_{cj}^T X_c u_c + \lambda_2 v_{cj}^T L_{vc} + \gamma_2' Y_{cj}^T Y_{cj} v_{cj}], \beta_2) \tag{11}$$

H is the soft-thresholding operator defined by $H(x, \beta) = sgn(x)\ max(|x| - \beta, 0)$ *and* $\gamma_2' = \gamma_2 + 1$.

After getting v, we can modify it to solve u. The objective function of u can be rewritten as follows.

$$\begin{aligned} &\min_{u,v_j} \sum_{c=1}^{n} \sum_{j=1}^{2} -u_c^T X_c^T Y_{cj} v_{cj} + \sum_{c=1}^{n} \frac{1}{n} \|z_c - X_c u_c\| \\ &+\beta_1 \|u_c\|_1 + \lambda_1 u_c^T L_{uc} u_c + \gamma_1 \|X_c u_c\|_2^2 \\ &s.t.\ \|X_c u_c\|_2^2 = 1, \|u_c\|_1 = d_3 \end{aligned} \tag{12}$$

In this way, we obtain a u value, which is the same as the method for (12), and we get the coordinate solution of (13)

$$\begin{aligned} u_c = H(\sum_{c=1}^{n} \sum_{j=1}^{2} [&X_c^T Y_{cj} v_{cj} + X_c^T z + \lambda_1 u_c^T L_{uc} \\ &+\gamma_1' X_c^T X u_c], \beta_1) \end{aligned} \tag{13}$$

Among them, $\gamma_1' = \gamma_1 + 1$. Now that the building blocks regarding updating u and each individual vj are created, we present the pseudocode in Algorithm 1.

III. RESULT

A. DATA SOURCE AND PREPROCESSING

In this section, we evaluate the effectiveness of the proposed method on the ADNI database (http://ADNI.loni.usc.edu/). In this database, we examine and select candidate SNP to predict the MRI imaging phenotype Reaction. In this study, there were 386 non-Hispanic Caucasian participants with

TABLE 1. Specific process for improved MTSCCA algorithm.

Algorithm for Improved MTSCCA
Input: Normalized data $\boldsymbol{X} \in R^{n\times p}$, $\boldsymbol{Y_j} \in R^{n\times p}$, and parameters λ_1、λ_2、β_1、β_2、γ_1、γ_2
Output: Canonical vectors v and u.
1: Initialization of $\boldsymbol{U} \in R^{p\times c}$ and $\boldsymbol{V} \in R^{q\times c}$
2: while no convergence do
3: for $c = 1$ *to* n do
4: for $j = 1$ to 2 do
5: for $i = 1$ to q do
6: $v_{cji} = H(\sum_{c=1}^{n}\sum_{j=1}^{2}[Y_{icj}^T X_{ic} u_{ic} + \lambda_2 v_{icj}^T L_{vic} + \gamma_2' Y_{icj}^T Y_{icj} v_{cji}], \beta_2)$
7: end for
8: end for
9: $V = v_c/\|v_c\|_1$
10: for i=1 to p do
11 $u_{ci} = H(\sum_{c=1}^{n}\sum_{j=1}^{2}[X_{ic}^T Y_{icj} v_{icj} + \frac{1}{n} X_{ic}^T z + \lambda_1 {u_{ic}}^T L_{uic} + \gamma_1' X_{ic}^T X u_{ci}], \beta_1)$
12: end for
13: $U = u_c/\|u_c\|_1$
14: end for
15: end while

imaging and genotyping data, including 113 healthy controls (HC), 248 mild cognitive impairment (MCI), and 25 AD Participants.

The original MRI images were downloaded from ADNI1 in the experiment, which used DiffusionKit [14] software to achieve head movement correction of the original MRI image. They registered it to the Montreal Neurological Institute (MNI) standard space. Next, the segmentation of MRI images was implemented by using MATLAB software CAT toolkit in the SPM software package [16]. Specifically, voxel-based morphometry (VBM) provides voxel estimation of the local number or volume of specific tissue compartments. By scaling the volume change due to spatial registration to adjust the segmentation, the volume of gray matter tissue was calculated in the region of interest (ROI) as a feature. After screening, in the end, 140 ROI were retained.

The genotypes of 386 subjects in this study came from the ADNI1 database. All SNPs were genotyped by the human 610 Quad BeadChip in the study. We used the genetic analysis tool PLINK [15] to screen genotype data, using the following exclusion criteria: rare SNPs (minor allele frequency (MAF) < 0.05), violations of Hardy-Weinberg Equilibrium

TABLE 2. Characteristics of the subjects.

Groups	AD	MCI	HC
Number	25	248	113
Gender(M/F)	10/15	133/115	58/55
Age(mean±std)	75.99±10.22	71.9±7.34	75.06±5.68
MMSE(mean)	20.48	27.90	29

Note: HC=Healthy Control, MCI=Mild Cognitive Impairment, AD=Alzheimer's disease.

(HWE p $< 10-6$), poor call rate ($< 90\%$) per subject and per SNP marker, gender check, and sibling pair identification.

We used ANNOVAR (http://ANNOVAR.openbioinformiscs.org) to annotate SNP with the corresponding genes listed in reference [17]. In order to get more accurate correlation analysis results, we focused on 15 genes. For each gene, we extracted all SNP within ± 5k base pairs of the gene boundary based on ANNOVAR annotation. Therefore, there is no overlap between the different groups. We selected 1000G phase3 v5 as the reference panel for genotype filling and used liftover software (https://genome.sph.umich.edu/wiki/LiftOver) to convert hg18-based SNP data to hg19-based data for better Compare with a reference panel. Then, the direction of the chain is corrected. The 1000G data is expressed based on the positive chain of the reference genome. To match the reference panel, snpflip is used to mark the positive chain, negative chain, and fuzzy chain in the sample data, and the blur is removed. Chain flipped the negative chain. Finally, the genotype filling was performed, first phasing through shapeit2 [18], [19], then through the Michigan Imputation Server website, the SNP data was filled. After another quality control, this resulted in 2957 single nucleotides being mapped to the 15 genes with the highest risk. Figure 1 shows the AD risk factor genes and the number of pre-selected SNPs in our study.

In Figure 1, the abscissa represents the name of the selected risk gene, and the ordinate represents the number of genes they contain. For the gene expression data corresponding to the sample, the repeated expression items were first deleted, and genes with variance more significant than 0.5 were extracted. Due to algorithm limitations, the gene expression data with the same number of SNPs was finally retained.

B. PARAMETER SELECTION

There are six parameters we need to set, namely β_1, β_2, γ_1, γ_2, λ_1, λ_2. Since blind grid search is very time consuming, we have adopted some techniques here to speed up the adjustment process. On the one hand, if the parameters are too small, SCCA and CCA will produce similar results. On the other hand, if the parameter is too large, SCCA will over-punish the result. Therefore, it is more reasonable for a parameter that is neither too large nor too small. We use the nested 5-fold cross-validation method to find the optimal parameters. These parameters generating the

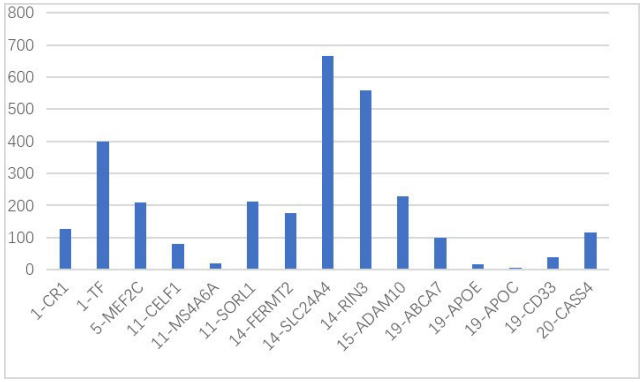

FIGURE 1. The number of 15 AD risk genes and their single nucleotide polymorphisms used in this study.

highest cross-task mean testing correlation coefficient will be chosen as the optimal parameters, i.e. $CV(\lambda, \beta, \gamma) = \frac{1}{5}\sum_{t=1}^{5}\sum_{j=1}^{2} Corr(\bar{X}_t u_t, \bar{Y}_{t,j} v_{t,j})$, where $\bar{X}_t$ and $\bar{Y}_{t,j}$ are the t-th fold which are the complementary sets used for testing, and u_t and $v_{t,j}$ are the canonical weights estimated from training set.

We fixed γ_1 =1 and γ_1 =1, because they mainly affect the amplitude of **U** and **V [20]**. Finally, we take $\beta_1 = \beta_2 = 10$, $\gamma_1 = \gamma_2 = 0.01$

C. IMPROVED BI-MULTIVARIATE ASSOCIATION

In this section, we will identify two sets of genetic data (SNP, Gene Expression Data) and a set of imaging phenotypes in the three categories of AD, MCI, and HC. For the three SCCA tasks, the proposed algorithm learns them together and generates an MRI weight matrix U, SNP weight matrix V1, and gene expression data weight matrix V2.

Figure 2 shows the weight of each QTs after five-fold cross-validation in the three data sets. Table 3 and Table 4 show the top 30 SNPs and the top 30 genes of the AD group and MCI group selected by our algorithm after five-fold cross-validation, and their respective absolute weights.

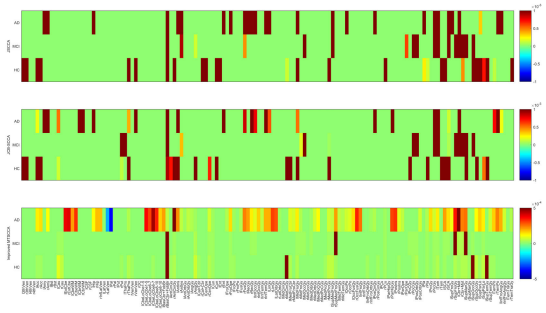

FIGURE 2. The weight of each QTs after five-fold cross-validation in the three sets of data.

IV. DISCUSSION

A. BIOLOGICAL SIGNIFICANCE

We observed that the non-zero coefficients of all these imaging markers have been shown to be significantly related to

TABLE 3. Use our model to select the genetic characteristics of the AD group.

SNPs	Weight	Gene	Weight
rs2362108	1.85E-04	GCLC	8.51E-05
rs143579966	1.84E-04	GCLM	7.75E-05
rs72693197	1.84E-04	LGALS2	7.25E-05
rs12434183	1.84E-04	LAMC1	7.22E-05
rs12432913	1.84E-04	CLIC4	7.18E-05
rs6494029	1.84E-04	RNAS3	7.15E-05
rs2362111	1.84E-04	RASGR3	7.13E-05
rs77247608	1.84E-04	PIK3AP1	7.05E-05
rs72697285	1.84E-04	E2F5	7.02E-05
rs11762396	1.84E-04	ATP1B3	7.02E-05
rs12433867	1.84E-04	RAD50	7.01E-05
rs11160077	1.84E-04	C8ORF9	6.95E-05
rs10135448	1.83E-04	IDH3A	6.94E-05
rs4900112	1.83E-04	LPP	6.93E-05
rs8177197	1.83E-04	RAB13	6.89E-05
rs58006834	1.83E-04	GFM1	6.87E-05
rs10134607	1.83E-04	DENNA	6.85E-05
rs74753154	1.83E-04	BPI	6.84E-05
rs7149739	1.83E-04	RPS4Y2	6.83E-05
rs4904927	1.77E-04	CCR6	6.82E-05
rs304138	1.77E-04	LTB4R	6.79E-05
rs1075472	1.77E-04	RNAS2	6.74E-05
rs9429780	1.77E-04	HAUS6	6.73E-05
rs9429941	1.77E-04	ISOC1	6.71E-05
rs12884739	0.000175	MS4A1	6.70E-05
rs10134595	1.75E-04	CAMP	6.60E-05
rs7159046	1.72E-04	NPAT	6.59E-05
rs2279796	1.67E-04	DNAJC5	6.59E-05
rs941639	1.66E-04	AMIG2	6.57E-05
rs67910622	1.65E-04	DEFA4	6.56E-05

the progression of AD. In addition to observing a slight difference in the gray matter volume of left/right hippocampus between the AD group and the MCI group and the HC group, significant differences in the gray matter volume of left/right precuneus and left parahippocampus between the two groups were also observed. These QTs have been reported in the past. It was confirmed in [21]–[23]. The JSCCA algorithm did not recognize the difference between precuneus and parahippocampus in the three groups. In addition, our algorithm also recognizes the weight difference between the three groups of

TABLE 4. Use our model to select the genetic characteristics of the MCI group.

SNPs	Weight	Gene	Weight
rs957191	0.022741	EIF1AX	8.51E-05
rs442495	0.011412	PGRMC1	7.75E-05
rs653765	0.011297	RPS4Y2	7.25E-05
rs12439457	0.011263	KDM5D	7.22E-05
rs347117	1.40E-05	RPS4Y1	7.18E-05
rs605928	1.40E-05	EIF1AY	7.15E-05
rs6494037	1.40E-05	PRKY	7.13E-05
rs11854073	1.40E-05	DDX3Y	7.05E-05
rs4775087	1.40E-05	LRP1	7.02E-05
rs7161889	1.40E-05	NELL2	7.02E-05
rs7163162	1.40E-05	AFF3	7.01E-05
rs2555356	1.40E-05	E2F5	6.95E-05
rs383902	1.40E-05	HIVEP2	6.94E-05
rs12911961	1.40E-05	NKRF	6.93E-05
rs12912003	1.36E-05	CAMP	6.89E-05
rs6494035	1.31E-05	LOC101927707	6.87E-05
rs6494036	1.31E-05	CNST	6.85E-05
rs7159655	1.31E-05	NT5E	6.84E-05
rs10138616	1.31E-05	MKL2	6.83E-05
rs11624016	1.31E-05	PRKX	6.82E-05
rs2146502	1.31E-05	LAMC1	6.79E-05
rs12434859	1.21E-05	FGD4	6.74E-05
rs700594	1.21E-05	BIRC3	6.73E-05
rs2146501	1.15E-05	FCRL1	6.71E-05
rs7157728	1.09E-05	RAB13	6.70E-05
rs12886904	1.09E-05	MMD	6.60E-05
rs700593	1.09E-05	BACH2	6.59E-05
rs7151162	1.07E-05	CR2	6.59E-05
rs7142157	1.07E-05	DHRSX	6.57E-05
rs4901316	7.98E-07	RCN3	6.56E-05

left/right posterior cingulate, which is the obvious difference between AD and MCI [24], [25]. Although JCB-SCCA uses the Laplacian matrix as a priori knowledge to introduce the algorithm, it also explains the biological significance of the selected QTs. Still, it does not accurately reflect the difference between these QTs in the diagnosis group and the control group. The results show that our proposed algorithm can identify the characteristics specific to the diagnosis group and reveal the biological significance of more QTs specific to the diagnosis group.

Also, we performed KEGG enrichment analysis on the TOP 600 genes in the AD group selected by the algorithm, as shown in figure 4A. Most of the biological processes are closely related to AD. Among them, the signalling pathways between B cells and T cells have been confirmed in previous literature [26]. Ginsenoside Rb1 can increase productions of Aβ1-42 and byproducts of β- and γ-secretase. Collected evidence supported that ginsenoside Rb1 improves learning and memory in AD rat by altering the amyloidogenic process of APP into the non-amyloidogenic process, to exert its anti-Inflammatory function [27]. In addition, RB1 selected by the algorithm as a tumor suppressor gene, its loss also has a carcinogenic risk [28]. Human ATP-binding cassette (ABC) transporters are potential causes or contributing factors of many pathologies such as AD. Literature [29] confirmed the potential of human ABC transporter as a new pharmacological target for the diagnosis and treatment of AD. The critical roles of apoptosis and hepatitis B in the process of AD were confirmed in the literature [30], [31]. Similarly, KEGG enrichment analysis was performed on the TOP 600 gene of the MCI group, as shown in Figure 4B. We found that the genes of the MCI group and the AD group are involved in many of the same biological processes. Cell adhesion molecules, hematopoietic cell lineage, and regulation of actin cytoskeleton are biological processes specific to the MCI group. In [32], amyloid β peptides (AβOs) have been shown to affect cognitive decline in AD and participate in cell adhesion molecules and regulation of actin cytoskeleton. The results indicate that AβOs plays a key role in the early stage of AD. The relationship between hematopoietic cell lineage and MCI needs to be further confirmed.

Not only that, the GCLC [33], GCLM [34], E2F5 [35], RAD50 [36], IDH3A [37], LPP [38], CCR6 [39], RNASE2 [40], ISOC1 [41], MS4A1 [42], CAMP [43] of the TOP 30 genes in the AD group selected by our proposed algorithm all play an important role in the AD process or participate in the biological processes related to it. In addition, in the MCI group, the low expression of PGRMC1 caused more than 90% of the exogenous Amyloid beta (Abeta) 1-42 oligomers to reduce the binding of neurons [44]. Abeta accumulates in the brains of patients with MCI. Low-density lipoprotein receptor-related protein 1 (LRP1) is involved in brain glucose metabolism and amyloid β clearance. It is shown in [45] that the T allele of LRP1 rs1799986 may reduce the sensitivity to MCI. Reference [46] indicates that phosphodiesterase (PDE) inhibitors can increase intracellular cyclic adenosine monophosphate (cAMP) to improve signal transduction pathways in brain circuits. The relationship between other genes and AD and MCI needs further confirmation.

Because the SNPs we selected are distributed near multiple risk genes, only rs2279796 can be confirmed in the existing literature [47] to have a strong correlation with AD. In the MCI group, rs442495, rs605928, rs7161889, and rs653765 have been confirmed in the [48], [49] to be SNP sites that are significantly related to MCI and AD near

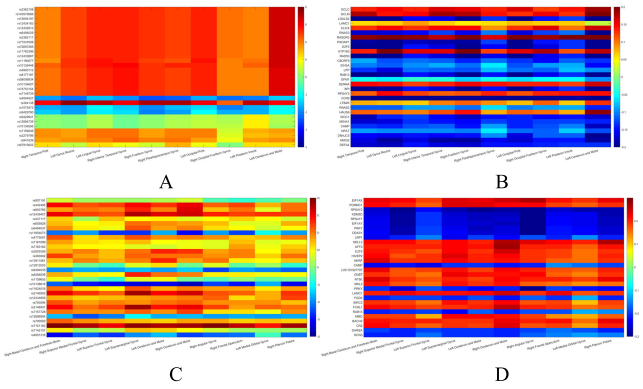

FIGURE 3. A and B are heat maps of the brain ROI-SNP and ROI-GENE associated with the selected markers at the top of the AD diagnosis group, respectively. C and D are heat maps of the brain ROI-SNP and ROI-GENE associated with the selected markers at the top of the MCI diagnosis group, respectively.

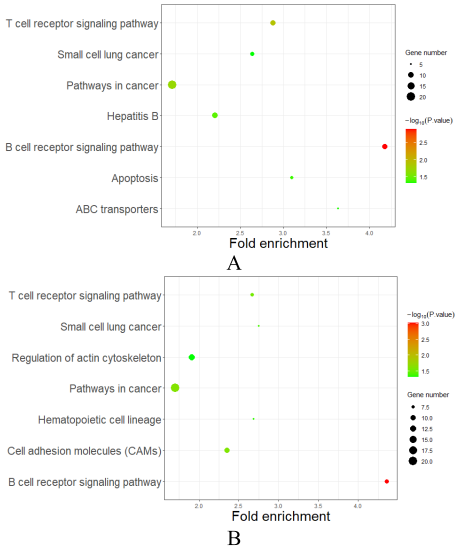

FIGURE 4. A and B respectively show the KEGG enrichment analysis results of the TOP 600 genes in the AD group and the MCI group selected by our algorithm.

ADAM10. And rs383902 [50] is a SNP site significantly related to schizophrenia.

In addition, based on the selected top 30 SNPs, top 30 gene and top 10 brain ROI, figure 3 shows the pair-wise correlation heat map of brain ROI-SNP/gene pairs. As expected, most ROI-SNP/gene pairs are strong. In addition, we also found that TOP 30 SNP and Genes corresponding to the same ROI showed a consistent positive/negative correlation. In order to find the significantly stronger SNP/gene-ROI pairs, we display the top five pairs with p $<$ 0.01 in the AD group and MCI group respectively in the Table 5 and Table 6. It can be seen from Table 5 that rs304138 is closely related to the changes in gray matter volume in multiple brain regions, and is likely to be a risk SNP site for Alzheimer's disease. It can be seen from Table 6 that Right Angular Gyrus are likely to be the risk brain areas of MCI, and its gray matter volume changes are jointly affected by multiple genes.

TABLE 5. The top five pairs with p<0.01 in the AD group.

SNP-ROI pairs	p-value	Gene-ROI pairs	p-value
rs304138 - Left Gyrus Rectus	3.15E-10	DHRSX - Left Cerebrum and Motor	7.49E-10
rs304138 - Right Fusiform Gyrus	1.93E-09	RCN3 - Left Cerebrum and Motor	6.77E-08
rs304138 - Right Parahippocampus Gyrus	3.90E-09	DHRSX - Right Inferior Temporal Gyrus	1.23E-07
rs304138 - Left Posterior Insula	8.95E-09	DHRSX - Left Occipital Pole	1.44E-07
rs2146502 - Left Lingual Gyrus	1.01E-08	NT5E - Right Fusiform Gyrus	1.90E-07

TABLE 6. The top five pairs with p $<$ 0.01 in the MCI group.

SNP-ROI pairs	p-value	Gene-ROI pairs	p-value
rs7151162 - Right Angular Gyrus	7.94E-13	EIF1AY - Right Angular Gyrus	3.50E-31
rs7151162 - Right Basal Cerebrum and Forebrain Brain	2.45E-11	KDM5D - Right Angular Gyrus	3.52E-31
rs7151162 - Left Supramarginal Gyrus	1.28E-10	RPS4Y1 - Right Angular Gyrus	4.26E-31
rs7142157 - Left Supramarginal Gyrus	5.00E-10	PRKY - Right Angular Gyrus	1.30E-29
rs2146501 - Right Planum Polare	5.41E-10	RPS4Y2 - Right Angular Gyrus	1.36E-29

B. COMPARISON WITH OTHER ALGORITHMS

In this research, we added the traditional SCCA algorithm to the multi-task framework. We used the Laplacian matrix as a new penalty item to process the image and genetic data. It can not only improve stability and anti-noise performance but also provide biologically interpretable results. The datasets are generated as follows. We first conducted experiments on the simulated data set. We simulated 100 random samples of neuroimaging data $\mathbf{X}$, $\mathbf{Y1}$(SNP) and $\mathbf{Y2}$(gene expression data) and applied the model similar to previous studies [17]. Then, define a latent variable ε, which belongs to the normal distribution $N(0, \sigma_\varepsilon^2)$. Finally, generate a neuroimaging loading vector α with p elements and a genetic loading vector β with q elements. Each non-zero elements was drawn independently from a uniform distribution $U(-1, -0.5) \cup U(0.5, 1)$.

We compare our method with the other three CCA-based multivariate methods (JCB-SCCA and JSCCA) and the MT-SCCALR algorithm that uses a multitasking framework. Our goal is to evaluate the correlation of SNP features, gene expression data features and MRI features in the simulation data set. We assume the imaging $\mathbf{X}$ and genotype

data **Y1** and **Y2**. The data set uses the same sample size and the number of features ($n = 300$, $p = 650$ and $q1 = q2 = 350$). Among them, the first data set mainly simulates task-specific situations, while the second data set mainly simulates tasks consistently. Specifically, in the second data set, there is a continuous relationship between the classes to affect the relationship between HC and MCI and between MCI and AD. The two data sets are generated as follows (Noise Level = 1). To evaluate the performance of these three algorithms. It can be seen from figure 5 that the anti-noise performance of the MR-SCCALR algorithm is poor. Other algorithms can better restore Ground Truth.

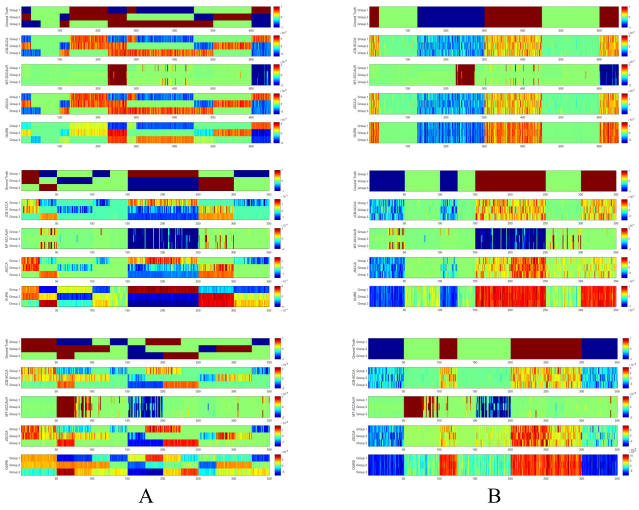

FIGURE 5. A and B are the norm weights of the two sets of data, respectively. Lines 1-5 are Ground truth, JCB-SCCA, MT-SCCALR, JSCCA and our proposed algorithm. For each data, from top to bottom are the weights U, V1, V2. In each panel, there are three rows (each row containing five times the norm weight) corresponding to three tasks.

The typical correlation coefficients CCC1 of **U** and **V1** and the typical correlation coefficients CCC2 of **U** and **V2** of the three sets of data under different noise levels. As shown in figure 6, we compared their anti-noise performance on three random data sets. These four algorithms exhibit relatively similar typical correlation coefficients when the noise is small. As the noise increases, our proposed algorithm shows better anti-noise performance than the other three algorithms.

We apply the proposed method with JCB-SCCA and JSCCA to this real neuroimaging genetic data. Due to the use of a large number of SNPs and genes, the MT-SCCALR method exceeds the program limit, and it is not added for comparison. All algorithms have three columns, representing AD, MCI, and HC. Figure 7 shows the typical correlation coefficients of the test set after five times of cross-validation of QTs-SNPs (U-V1) and QTs-Gene Expression Data (U-V2).

We show the test classification accuracy of SVM based on the LIBSVM (https://www.csie.ntu.edu.tw/∼cjlin/libsvm/) software package respectively. In Table 7, our proposed algorithm has the highest classification accuracy in the AD

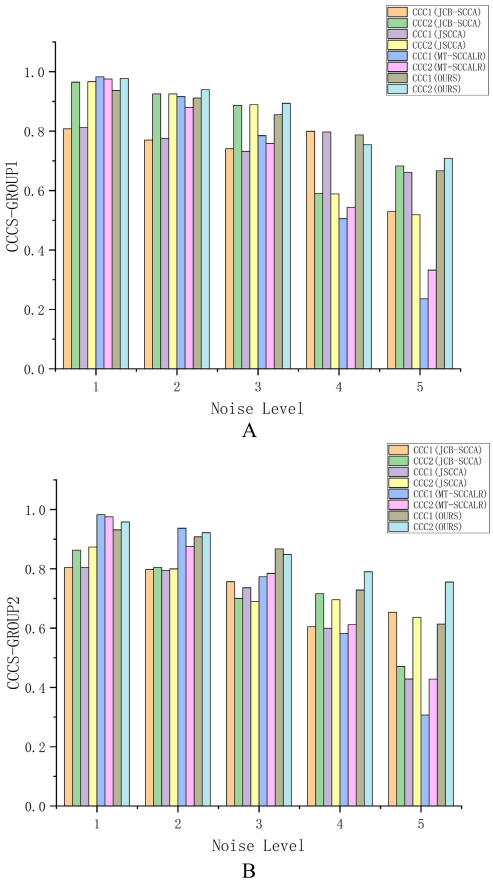

FIGURE 6. The four algorithms are compared with five-fold cross-validation under different noise levels. A and B are the sets of simulation data, respectively.

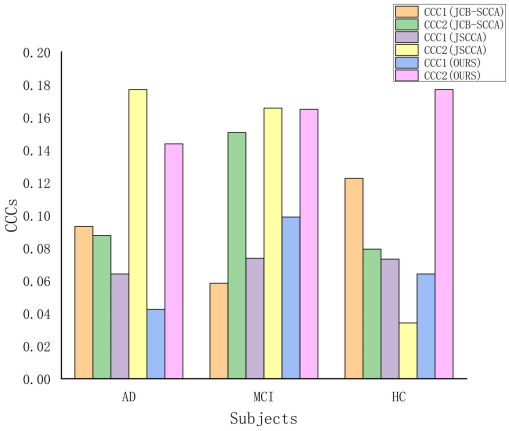

FIGURE 7. CCC comparison of three algorithms with five-fold cross validation on different data sets.

and HC data sets and achieved the same classification accuracy in the MCI data set with JCB-SCCA. In addition, we found that the JSCCA and JCB-SCCA algorithms show higher classification accuracy on the training set than our algorithm, but they perform poorly on the test set. It indicates that our proposed algorithm can more effectively prevent

TABLE 7. Classification accuracy using SVM algorithm.

	Train			Test		
	AD	MCI	HC	AD	**MCI**	HC
JSCCA	**0.9377**	**0.6684**	**0.7388**	0.8084	0.6372	0.5440
JCB-SCCA	0.9330	0.6632	0.7225	0.8034	**0.6452**	0.5211
Improved MTSCCA	0.9132	0.6464	0.7087	**0.8213**	0.6373	**0.5649**

over-simulation. It reflects the robustness of our proposed algorithm to a certain extent.

V. CONCLUSION

SCCA is a robust and scalable multiple association analysis algorithm, which has been widely used in the field of image genetics [2]–[9]. Compared with the constraints of existing methods, we propose an improved multi-task SCCA method. It uses one kind of image (sMRI) and two kinds of genetic data (SNP, gene expression data) to mine the characteristics specific to each diagnosis group. In addition, the Laplacian matrix is used as the algorithm penalty item to make the results easier to interpret and have biological significance. We tested our algorithm on the simulation data set and the ADNI data set. The results show that, compared with the existing methods (JCB-SCCA, JSCCA, MT-SCCALR), the algorithm not only has stronger anti-noise performance in the simulation data set, but also has higher canonical correlation coefficient and classification accuracy in the diagnosis group of real data set. The proposed algorithm can mine the unique characteristics of different diagnostic groups, and also found many SNP-ROI and gene-ROI pairs that may be closely related to MCI and AD, which are worthy of in-depth study.

In addition, we used one type of imaging data (MRI) and two types of genetic data (SNP, Gene Expression Data). In future research, we will try to integrate more imaging modalities (such as PET, fMRI), and more types of genetic data (such as DNA methylation data), hoping to more accurately and comprehensively explore biological significance between the brain area and the risk gene locus.

ACKNOWLEDGMENT

(Kai Wei and Wei Kong are co-first authors.)

REFERENCES

[1] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, Dec. 1936.

[2] E. Parkhomenko, D. Tritchler, and J. Beyene, "Sparse canonical correlation analysis with application to genomic data integration," *Stat. Appl. Genet. Mol. Biol.*, vol. 8, no. 1, pp. 1–34, Jan. 2009, doi: 10.2202/1544-6115.1406.

[3] L. Du, K. Liu, X. Yao, S. L. Risacher, J. Han, A. J. Saykin, L. Guo, and L. Shen, "Detecting genetic associations with brain imaging phenotypes in Alzheimer's disease via a novel structured SCCA approach," *Med. Image Anal.*, vol. 61, Apr. 2020, Art. no. 101656, doi: 10.1016/j.media.2020.101656.

[4] J. Yan, L. Du, S. Kim, S. L. Risacher, H. Huang, J. H. Moore, A. J. Saykin, L. Shen, and f. the Alzheimer's Disease neuroimaging Initiative, "Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm," *Bioinformatics*, vol. 30, no. 17, pp. i564–i571, Sep. 2014, doi: 10.1093/bioinformatics/btu465.

[5] L. Du, H. Huang, J. Yan, S. Kim, S. L. Risacher, M. Inlow, J. H. Moore, A. J. Saykin, and L. Shen, "Structured sparse canonical correlation analysis for brain imaging genetics: An improved GraphNet method," *Bioinformatics*, vol. 32, no. 10, pp. 1544–1551, May 2016, doi: 10.1093/bioinformatics/btw033.

[6] X. Hao, C. Li, J. Yan, X. Yao, S. L. Risacher, A. J. Saykin, L. Shen, and D. Zhang, "Identification of associations between genotypes and longitudinal phenotypes via temporally-constrained group sparse canonical correlation analysis," *Bioinformatics*, vol. 33, no. 14, pp. i341–i349, Jul. 2017, doi: 10.1093/bioinformatics/btx245.

[7] L. Du, K. Liu, X. Yao, S. L. Risacher, J. Han, A. J. Saykin, L. Guo, and L. Shen, "Multi-task sparse canonical correlation analysis with application to multi-modal brain imaging genetics," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 1, pp. 227–239, Feb. 2021, doi: 10.1109/TCBB.2019.2947428.

[8] L. Du, F. Liu, K. Liu, X. Yao, S. L. Risacher, J. Han, L. Guo, A. J. Saykin, and L. Shen, "Identifying diagnosis-specific genotype–phenotype associations via joint multitask sparse canonical correlation analysis and classification," *Bioinformatics*, vol. 36, no. 1, pp. i371–i379, Jul. 2020, doi: 10.1093/bioinformatics/btaa434.

[9] A. C. Lorena, A. C. P. L. F. de Carvalho, and J. M. P. Gama, "A review on the combination of binary classifiers in multiclass problems," *Artif. Intell. Rev.*, vol. 30, nos. 1–4, pp. 19–37, Dec. 2008.

[10] L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, and J. E. Taylor, "Interpretable whole-brain prediction analysis with GraphNet," *NeuroImage*, vol. 72, pp. 304–321, May 2013, doi: 10.1016/j.neuroimage.2012.12.062.

[11] M. Kim, J. H. Won, J. Youn, and H. Park, "Joint-connectivity-based sparse canonical correlation analysis of imaging genetics for detecting biomarkers of Parkinson's disease," *IEEE Trans. Med. Imag.*, vol. 39, no. 1, pp. 23–34, Jan. 2020, doi: 10.1109/TMI.2019.2918839.

[12] M. E. Levine, P. Langfelder, and S. Horvath, "A weighted SNP correlation network method for estimating polygenic risk scores," in *Biological Networks and Pathway Analysis*, vol. 1613,2017, pp. 277–290, doi: 10.1007/978-1-4939-7027-8_10.

[13] G. D. Pearlson, J. Liu, and V. D. Calhoun, "An introductory review of parallel independent component analysis (p-ICA) and a guide to applying p-ICA to genetic data and imaging phenotypes to identify disease-associated biological pathways and systems in common complex disorders," *Frontiers Genet.*, vol. 6, p. 276, Sep. 2015, doi: 10.3389/fgene.2015.00276.

[14] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: A survey and extensions," *Math. Methods Operations Res.*, vol. 66, no. 3, pp. 373–407, Nov. 2007.

[15] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, and P. C. Sham, "PLINK: A tool set for whole-genome association and population-based linkage analyses," *Amer. J. Hum. Genet.* vol. 81, vol. 3, pp. 559–575, 2007, doi: 10.1086/519795.

[16] A. J. Saykin, L. Shen, T. M. Foroud, S. G. Potkin, S. Swaminathan, S. Kim, S. L. Risacher, K. Nho, M. J. Huentelman, D. W. Craig, and P. M. Thompson, "Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans," *Alzheimer's Dementia, J. Alzheimer'S Assoc.*, vol. 6, no. 3, pp. 265–273, 2010, doi: 10.1016/j.jalz.2010.03.013.

[17] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Res.*, vol. 38, no. 16, p. e164, Sep. 2010, doi: 10.1093/nar/gkq603.

[18] O. Delaneau, J. Marchini, and J.-F. Zagury, "A linear complexity phasing method for thousands of genomes," *Nature Methods*, vol. 9, no. 2, pp. 179–181, Feb. 2012, doi: 10.1038/nmeth.1785.

[19] J. O'Connell *et al.*, "A general approach for haplotype phasing across the full spectrum of relatedness," *PLoS Genet.*, vol. 10, no. 4, Apr. 2014, Art. no. e1004234, doi: 10.1371/journal.pgen.1004234.

[20] J. Chen, F. D. Bushman, J. D. Lewis, G. D. Wu, and H. Li, "Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis," *Biostatistics*, vol. 14, no. 2, pp. 244–258, Apr. 2013, doi: 10.1093/biostatistics/kxs038.

[21] A. Fellgiebel and I. Yakushev, "Diffusion tensor imaging of the hippocampus in MCI and early Alzheimer's disease," *J. Alzheimer's Disease*, vol. 26, no. 3, pp. 257–262, 2011, doi: 10.3233/JAD-2011-0001.

[22] G. Aghakhanyan, A. Vergallo, M. Gennaro, S. Mazzarri, F. Guidoccio, C. Radicchi, R. Ceravolo, G. Tognoni, U. Bonuccelli, and D. Volterrani, "The precuneus—A witness for excessive A? Gathering in Alzheimer's disease pathology," *Neurodegenerative Diseases*, vol. 18, nos. 5–6, pp. 302–309, 2018, doi: 10.1159/000492945.

[23] L. Delano-Wood, N. H. Stricker, S. F. Sorg, D. A. Nation, A. J. Jak, S. P. Woods, D. J. Libon, D. C. Delis, L. R. Frank, and M. W. Bondi, "Posterior cingulum white matter disruption and its associations with verbal memory and stroke risk in mild cognitive impairment," *J. Alzheimer's Disease*, vol. 29, no. 3, pp. 589–603, Mar. 2012, doi: 10.3233/JAD-2012-102103.

[24] G. B. Frisoni, R. Ganzola, E. Canu, U. Rüb, F. B. Pizzini, F. Alessandrini, G. Zoccatelli, A. Beltramello, C. Caltagirone, and P. M. Thompson, "Mapping local hippocampal changes in Alzheimer's disease and normal ageing with MRI at 3 tesla," *Brain, J. Neurology* vol. 131, no. 12, pp. 3266–3276, 2008, doi: 10.1093/brain/awn280.

[25] J. Keaney, J. Gasser, G. Gillet, D. Scholz, and I. Kadiu, "Inhibition of Bruton's tyrosine kinase modulates microglial phagocytosis: Therapeutic implications for Alzheimer's disease," *J. Neuroimmune Pharmacol.*, vol. 14, no. 3, pp. 448–461, Sep. 2019, doi: 10.1007/s11481-019-09839-0.

[26] K. Schindowski, A. Eckert, J. Peters, C. Gorriz, U. Schramm, T. Weinandi, K. Maurer, L. Frölich, and W. E. Müller, "Increased T-cell reactivity and elevated levels of CD8+ memory T-cells in Alzheimer's disease-patients and T-cell hyporeactivity in an Alzheimer's disease-mouse model: Implications for immunotherapy," *Neuromol. Med.*, vol. 9, vol. 4, pp. 340–354, 2007, doi: 10.1007/s12017-007-8015-9.

[27] J. Lin, S. Gao, T. Wang, Y. Shen, W. Yang, Y. Li, and H. Hu, "Ginsenoside Rb1 improves learning and memory ability through its anti-inflammatory effect in Aβ1-40 induced Alzheimer's disease of rats," *Amer. J. Transl. Res.* vol. 11, pp. 2955–2968, May 2019.

[28] G. Mollaoglu *et al.*, "MYC drives progression of small cell lung cancer to a variant neuroendocrine subtype with vulnerability to aurora kinase inhibition," *Cancer Cell*, vol. 31, no. 2, pp. 270–285, Feb. 2017, doi: 10.1016/j.ccell.2016.12.005.

[29] C. D. Pereira, F. Martins, J. Wiltfang, O. A. B. da Cruz e Silva, and S. Rebelo, "ABC transporters are key players in Alzheimer's disease," *J. Alzheimer's Disease*, vol. 61, no. 2, pp. 463–485, Dec. 2017, doi: 10.3233/JAD-170639.

[30] M. Obulesu and M. J. Lakshmi, "Apoptosis in Alzheimer's disease: An understanding of the physiology, pathology and therapeutic avenues," *Neurochem. Res.*, vol. 39, no. 12, pp. 2301–2312, 2014, doi: 10.1007/s11064-014-1454-4.

[31] D. Mastroeni, J. Nolz, S. Sekar, E. Delvaux, G. Serrano, L. Cuyugan, W. S. Liang, T. G. Beach, J. Rogers, and P. D. Coleman, "Laser-captured microglia in the Alzheimer's and Parkinson's brain reveal unique regional expression profiles and suggest a potential role for hepatitis b in the Alzheimer's brain," *Neurobiol. Aging*, vol. 63, pp. 12–21, Mar. 2018, doi: 10.1016/j.neurobiolaging.2017.10.019.

[32] A. Sebollela, L. Freitas-Correa, F. F. Oliveira, A. C. Paula-Lima, L. M. Saraiva, S. M. Martins, L. D. Mota, C. Torres, S. Alves-Leon, J. M. De Souza, and D. M. Carraro, "Amyloid-β oligomers induce differential gene expression in adult human brain slices," *J. Biol. Chem.* vol. 287, no. 10, pp. 7436–7445, 2012, doi: 10.1074/jbc.M111.298471.

[33] S. Lee, Y. Chiu, S. Yang, C. Chen, C. Huang, G. Lee-Chen, W. Lin, and K. Chang, "Novel synthetic chalcone-coumarin hybrid for Aβ aggregation reduction, antioxidation, and neuroprotection," *CNS Neurosci. Therapeutics*, vol. 24, no. 12, pp. 1286–1298, Dec. 2018, doi: 10.1111/cns.13058.

[34] L. Luo, S.-W. Kim, H.-K. Lee, I.-D. Kim, H. Lee, and J.-K. Lee, "Gastrodin exerts robust neuroprotection in the postischemic brain via its protective effect against Zn2+-toxicity and its anti-oxidative effects in astrocytes," *Animal Cells Syst.*, vol. 22, no. 6, pp. 429–437, Nov. 2018, doi: 10.1080/19768354.2018.1549099.

[35] C. E. Johanson, J. A. Duncan, P. M. Klinge, T. Brinker, E. G. Stopa, and G. D. Silverberg, "Multiplicity of cerebrospinal fluid functions: New challenges in health and disease," *Cerebrospinal Fluid Res.*, vol. 5, no. 1, pp. 1–32, May 2008, doi: 10.1186/1743-8454-5-10.

[36] E. Jacobsen, T. Beach, Y. Shen, R. Li, and Y. Chang, "Deficiency of the Mre11 DNA repair complex in Alzheimer's disease brains," *Mol. Brain Res.*, vol. 128, no. 1, pp. 1–7, Sep. 2004, doi: 10.1016/j.molbrainres.2004.05.023.

[37] A. Fiorini, R. Sultana, S. Förster, M. Perluigi, G. Cenini, C. Cini, J. Cai, J. B. Klein, S. A. Farr, M. L. Niehoff, J. E. Morley, V. B. Kumar, and D. A. Butterfield, "Antisense directed against PS-1 gene decreases brain oxidative markers in aged senescence accelerated mice (SAMP8) and reverses learning and memory impairment: A proteomics study," *Free Radical Biol. Med.*, vol. 65, pp. 1–14, Dec. 2013, doi: 10.1016/j.freeradbiomed.2013.06.017.

[38] E. Elkind, T. Vaisid, J. D. Kornspan, S. Barnoy, S. Rottem, and N. S. Kosower, "Calpastatin upregulation in Mycoplasma hyorhinis-infected cells is promoted by the mycoplasma lipoproteins via the NF-κB pathway," *Cellular Microbiol.* vol. 14, no. 6, pp. 840–851, 2012, doi: 10.1111/j.1462-5822.2012.01760.x.

[39] S. Subramanian, P. Ayala, T. L. Wadsworth, C. J. Harris, A. A. Vandenbark, J. F. Quinn, and H. Offner, "CCR6: A biomarker for Alzheimer's-like disease in a triple transgenic mouse model," *J. Alzheimer's Disease*, vol. 22, no. 2, pp. 619–629, 2010, doi: 10.3233/JAD-2010-100852.

[40] S. Mukherjee, S. Kim, L. E. Gibbons, K. Nho, S. L. Risacher, M. M. Glymour, C. Habeck, G. J. Lee, E. Mormino, N. Ertekin-Taner, and T. J. Montine, "Genetic architecture of resilience of executive functioning," *Brain Imag. Behav.* vol. 6, vol. 4, pp. 621–633, 2012, doi: 10.1007/s11682-012-9184-1.

[41] S. Jiang, L. Tang, N. Zhao, W. Yang, Y. Qiu, and H.-Z. Chen, "A systems view of the differences between APOE ε4 carriers and non-carriers in Alzheimer's disease," *Frontiers Aging Neurosci.*, vol. 8, p. 171, Jul. 2016, doi: 10.3389/fnagi.2016.00171.

[42] J. Ma, J. T. Yu, and L. Tan, "MS4A cluster in Alzheimer's disease," *Mol. Neurobiol.*, vol. 51, no. 3, pp. 1240–1248, 2015, doi: 10.1007/s12035-014-8800-z.

[43] S. M. Nabavi, S. Talarek, J. Listos, S. F. Nabavi, K. P. Devi, M. R. de Oliveira, D. Tewari, S. Argüelles, S. Mehrzadi, A. Hosseinzadeh, G. D'onofrio, I. E. Orhan, A. Sureda, S. Xu, S. Momtaz, and M. H. Farzaei, "Phosphodiesterase inhibitors say NO to Alzheimer's disease," *Food Chem. Toxicology*, vol. 134, Dec. 2019, Art. no. 110822, doi: 10.1016/j.fct.2019.110822.

[44] N. J. Izzo *et al.*, "Alzheimer's therapeutics targeting amyloid beta 1–42 oligomers II: Sigma-2/PGRMC1 receptors mediate abeta 42 oligomer binding and synaptotoxicity," *PLoS ONE*, vol. 9, no. 11, Nov. 2014, Art. no. e111899, doi: 10.1371/journal.pone.0111899.

[45] W. Cao, S. Tian, H. Zhang, W. Zhu, K. An, J. Shi, Y. Yuan, and S. Wang, "Association of low-density lipoprotein receptor-related protein 1 and its rs1799986 polymorphism with mild cognitive impairment in chinese patients with type 2 diabetes," *Frontiers Neurosci.*, vol. 14, p. 743, Sep. 2020, doi: 10.3389/fnins.2020.00743.

[46] J. Prickaerts, P. R. A. Heckman, and A. Blokland, "Investigational phosphodiesterase inhibitors in phase i and phase II clinical trials for Alzheimer's disease," *Expert Opinion Investigational Drugs*, vol. 26, no. 9, pp. 1033–1048, Sep. 2017, doi: 10.1080/13543784.2017.1364360.

[47] J. B. Vasquez, D. W. Fardo, and S. Estus, "ABCA7 expression is associated with Alzheimer's disease polymorphism and disease status," *Neurosci. Lett.*, vol. 556, pp. 58–62, Nov. 2013, doi: 10.1016/j.neulet.2013.09.058.

[48] M. Kim, J. Suh, D. Romano, M. H. Truong, K. Mullin, B. Hooli, D. Norton, G. Tesco, K. Elliott, S. L. Wagner, and R. D. Moir, "Potential late-onset Alzheimer's disease-associated mutations in the ADAM10 gene attenuate alpha-secretase activity," *Hum. Mol. Genet.* vol. 18, no. 20, pp. 3987–3996, 2009, doi: 10.1093/hmg/ddp323.

[49] F. Zeng, C. Shen, Y. H. Liu, J. Li, J. Zhu, Y. R. Wang, J. C. Yan, C. Y. Gao, H. D. Zhou, J. Deng, and Y. J. Wang, "Genetic association between APP, ADAM10 gene polymorphism, and sporadic Alzheimer's disease in the Chinese population," *Neurotoxicity Res.*, vol. 27, vol. 3, pp. 284–291, 2015, doi: 10.1007/s12640-015-9516-1.

[50] P. A. Holmans, B. Riley, A. E. Pulver, M. J. Owen, D. B. Wildenauer, P. V. Gejman, B. J. Mowry, C. Laurent, K. S. Kendler, G. Nestadt, and N. M. Williams, "Genomewide linkage scan of schizophrenia in a large multicenter pedigree sample using single nucleotide polymorphisms," *Mol. Psychiatry*, vol. 14, no. 8, pp. 786–795, 2009, doi: 10.1038/mp.2009.11400(9):1181-1189.

KAI WEI received the B.S. degree from Shanghai Dianji University, China, in 2019. He is currently pursuing the master's degree with Shanghai Maritime University. His main research interest includes bioinformatics areas, including genomics data and imaging data analysis.

WEI KONG received the bachelor's and master's degrees in control theory and control engineering from the Qingdao University of Science and Technology, China, in 1998 and 2001 respectively, and the Ph.D. degree in the pattern recognition and intelligence system from Shanghai Jiao Tong University, China, in 2005. After her graduation, she joined the College of Information Engineering, Shanghai Maritime University, in 2015, where she served as a Professor for bioinformatics. From 2007 to 2008, she worked as a Visitor Scholar with the Biomedical Informatics Laboratory, Harvard Medical School, Harvard University, USA. Her current research interest includes bioinformatics areas, including genomics data analysis, pathway, and regulatory network reconstruction of dementia and cancers. She has been supervising many grants, including two National Natural Science Foundation of China (NSFC), one Natural Science Foundation of Shanghai, three Scientific Innovation Projects supported by the Shanghai Municipal Education Commission, one Foundation of Science and Technology Phosphor (A Plan) supported by the Science and Technology Commission of Shanghai Municipality, and grants of Shanghai Maritime University.

SHUAIQUN WANG received the B.S. degree from Inner Mongolia Normal University, Huhehaote, China, in 2007, the M.S. degree from Shanghai Maritime University, Shanghai, China, in 2010, and the D.E. degree from Tongji University, Shanghai, in 2015. She is currently working with Shanghai Maritime University. Her main research interest includes theoretical research and application of intelligent optimization algorithm, such as function optimization, fuzzy shop scheduling problem, and information feature extraction of biological data.