DEEP NETWORK-BASED FEATURE SELECTION FOR IMAGING GENETICS: APPLICATION TO IDENTIFYING BIOMARKERS FOR PARKINSON'S DISEASE

*Mansu Kim[1,2,4], Ji Hye Won[1,2], Jisu Hong[1,2], Junmo Kwon[1,2], Hyunjin Park[2,3], and Li Shen[4]**

[1] Department of Electrical and Computer Engineering, Sungkyunkwan University, Korea
[2] Center for Neuroscience Imaging Research, Institute for Basic Science, Korea
[3] School of Electronic and Electrical Engineering, Sungkyunkwan University, Korea
[4] Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, USA

ABSTRACT

Imaging genetics is a methodology for discovering associations between imaging and genetic variables. Many studies adopted sparse models such as sparse canonical correlation analysis (SCCA) for imaging genetics. These methods are limited to modeling the linear imaging genetics relationship and cannot capture the non-linear high-level relationship between the explored variables. Deep learning approaches are underexplored in imaging genetics, compared to their great successes in many other biomedical domains such as image segmentation and disease classification. In this work, we proposed a deep learning model to select genetic features that can explain the imaging features well. Our empirical study on simulated and real datasets demonstrated that our method outperformed the widely used SCCA method and was able to select important genetic features in a robust fashion. These promising results indicate our deep learning model has the potential to reveal new biomarkers to improve mechanistic understanding of the studied brain disorders.

Index Terms— Imaging genetics, feature selection, deep learning, Parkinson's disease

1. INTRODUCTION

Imaging genetics is an emerging research field, where the relationship between imaging and genetic variables is investigated to better understand genetic determinants of imaging phenotypes. Several machine learning approaches were proposed to identify important genetic and imaging features and reveal the association between them. Typical previous studies applied sparse canonical correlation analysis (SCCA), reduced-rank regression, and sparse regularized linear models to identify the association between imaging and genetic variables [1]–[3].

Parkinson's disease (PD) is the second most common neurodegenerative disorder which triggers various motor symptoms, such as bradykinesia, rigidity, resting tremor, and postural instability. Although the neurological features of PD are rather well-defined, such as the loss of dopaminergic neurons in the substantia nigra, the underlying genetic causal relationship of PD is unclear. Recently, various neuroimaging techniques have been employed to investigate the effects of PD on the brain. A genome-wide association study (GWAS) revealed common genetic variants related to PD. As an extension, integrating complementary imaging and genetic information to study PD has become an important topic.

Deep learning is a type of computational models with multiple processing layers. These approaches have made huge successes in solving problems from many domains. Specifically, the methods have dramatically improved performance in speech recognition, classification, and segmentation. Also, a few studies investigated the usage of deep learning for feature selection. For example, Li *et. al.,* proposed a deep feature selection (DFS) architecture that adopted one-to-one connected layer as the first hidden layer of the deep network [4]. This method did not support multi-modal data (e.g., imaging and genetics data), thus is not suitable for imaging genetics.

In this study, we proposed a deep network-based feature selection model for evaluating associations between genetic and neuroimaging data. Specifically, we analyzed genetic markers such as single nucleotide polymorphisms (SNPs) and neuroimaging measures extracted from dopamine transporter single photon emission computed tomography (DaT-SPECT) scans. Our proposed network model is a genetics-to-image circulation network. Our network includes two one-to-one connected layers with least absolute shrinkage and selection operator (LASSO) regularization to identify sparse significant features that associate genetic and

*Correspondence to Li Shen (Li.Shen@pennmedicine.upenn.edu).
This study was supported by the Institute for Basic Science (IBS-R015-D1), the National Research Foundation of Korea (NRF-2019R1H1A2079721), the IITP grant funded by the Korean government under the AI Graduate School Support Program (2019-0-00421), and the Ministry of Science and ICT of Korea under the ITRC program (grant number IITP-2019-2018-0-01798). The Imaging data were obtained from the PPMI. This work was also supported in part by the National Institutes of Health [R01 EB022574, RF1 AG063481] and National Science Foundation [IIS 1837964] at the University of Pennsylvania.

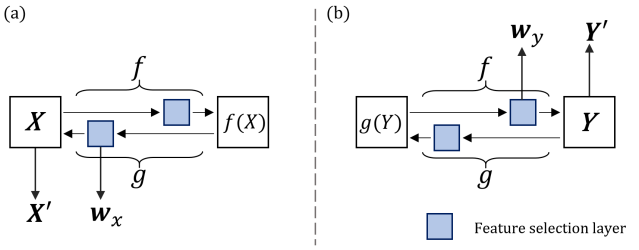

Fig. 1. The overview of our end-to-end deep network-based feature selection architecture. Subfigures (a) and (b) denote the detailed procedures to extract selection vectors of two different inputs, where $\boldsymbol{X}, \boldsymbol{Y}$, $\boldsymbol{X}'$, $\boldsymbol{Y}', \boldsymbol{w}_x$, and $\boldsymbol{w}_y$ denote SNP, DaT-SPECT, estimated SNP, estimated DaT-SPECT, selection weight vector of SNP, and selection weight vector of DaT-SPECT, respectively. The functions $\boldsymbol{f}$ and $\boldsymbol{g}$ denote a SNP-to-image mapping function and image-to-SNP mapping function, respectively.

neuroimaging data. Our contributions are as follows: i) A novel deep network-based feature selection method is proposed and applied to imaging genetics problems. ii) The proposed algorithm is an unsupervised learning model, and thus could be adopted in unlabeled data. iii) We have applied our proposed algorithm to both simulation and real data (i.e., PD patient) and compared the results with those obtained using the SCCA method. Our algorithm identified several well-established PD biomarkers and revealed new potential SNPs.

2. METHODS

2.1. Sparse canonical correlation analysis (SCCA)

Herein, we use the boldface lowercase letter to denote a vector, and the boldface uppercase letter to denote a matrix. Specifically, given datasets $\boldsymbol{X} \in \mathbb{R}^{n\times p}, \boldsymbol{Y} \in \mathbb{R}^{n\times q}$ with n samples, $\boldsymbol{X}$ denotes p features of the SNP data, and $\boldsymbol{Y}$ denotes q features of the neuroimaging data [5]. The sparse canonical correlation analysis (SCCA) method aims to identify the maximized correlation between two datasets as follows:

$$\min_{u,v} -\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v} + \lambda_1\|\boldsymbol{u}\|_1 + \lambda_2\|\boldsymbol{v}\|_1 \qquad (1)$$

where $\boldsymbol{u}$ and $\boldsymbol{v}$ are the corresponding selection vectors. The λ_1 and λ_2 denote regularization parameters of LASSO penalty which control the model sparsity.

2.2. Deep network-based feature selection (DN-FS)

2.2.1. Proposed Architecture

We assume that there exist significant cross-modal relations between genetic and imaging data (e.g., SNPs and DaT-SPECT) and our goal is to apply feature selection approach to discover such relations between two different data modalities. This is different from DFS that typically is supervised and hence requires paired input data and label. In this study, we propose a deep network-based feature selection (DN-FS) model for evaluating associations between genetic and neuroimaging data. Of note, our approach belongs to an unsupervised category.

The idea of circulation network comes from the image-to-image translation model, such as cycle generative adversarial network, proposed by Jun-Yan Zhu *et. al.* [6]. The network architecture consists of two circulation networks as shown in **Fig. 1**. The first circular network translates the input data $\boldsymbol{X}$ (i.e., SNPs) to another input data space (i.e. DaT-SPECT) using a deep neural network ($\boldsymbol{f}$) and then translates back again to estimate data $\boldsymbol{X}'$ with selection vector ($\boldsymbol{w}_x$) using a deep neural network ($\boldsymbol{g}$) (**Fig. 1-a**). In the second network, input data $\boldsymbol{Y}$ (i.e., DaT-SPECT) is mapped to another input data space using the network $\boldsymbol{f}$ and then translates back using the network $\boldsymbol{g}$ to estimate data $\boldsymbol{Y}'$ with selection vector ($\boldsymbol{w}_y$) (**Fig. 1-b**).

2.2.2. Feature selection with deep learning

We borrowed two ideas from conventional machine learning literature to select correlated features between two data sets. First, we added a one-to-one mapping layer, named feature selection layer, between the last hidden layer and the output layer (**Fig. 1**). The *i*-th node of the feature selection layer was only connected to the *i*-th output element with linear activation function. Then, we applied the LASSO penalty to force the selection vectors to be sparse. The second idea was to add a CCA penalty to the cost function since the mean squared error between $\boldsymbol{X}$ and $\boldsymbol{X}'$ or $\boldsymbol{Y}$ and $\boldsymbol{Y}'$ cannot guarantee that the select features from $\boldsymbol{X}$ and $\boldsymbol{Y}$ were significantly correlated. Thus, the proposed full objective function is as follow:

$$\min \|\boldsymbol{X}-\boldsymbol{X}'\|_F^2 + \|\boldsymbol{Y}-\boldsymbol{Y}'\|_F^2 - \boldsymbol{w}_x{}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{w}_y + \lambda_1\|\boldsymbol{w}_x\|_1 + \lambda_2\|\boldsymbol{w}_y\|_1 \qquad (2)$$

2.2.3. Implementation and training details

We adopted a shallow fully connected network as mapping function (i.e., $\boldsymbol{f}$ and $\boldsymbol{g}$). For the simulation setup, we used three hidden layers ({500, 2000, 100} neurons for SNP-to-image mapping and {100, 2000, 500} neurons for image-to-SNP mapping). For real imaging genetics setup, we also used three hidden layers ({3365, 2000, 90} neurons for SNP-to-image mapping and {90, 2000, 3365} neurons for image-to-SNP mapping). Each hidden layer includes a rectified linear unit as activation function followed by batch normalization. For all the experiments, we tuned the hyper-parameters, λ_1 and λ_2, using a grid search algorithm. Both parameters were jointly tuned by nested five-fold cross-validation to maximize the averaged correlation. We set $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$ for the proposed model and $\lambda_1 = 0.1$ and $\lambda_2 = 0.1$ for the SCCA model. We used adaptive gradient optimizer implemented in TensorFlow. The learning rate was 0.1 and a batch size was 200.

2.3. Neuroimaging genetics data processing

We obtained neuroimaging (i.e., DaT-SPECT) and genotyping data from the Parkinson's Progression Markers Initiative (PPMI) database. In detail, we used 94 PD cases of unrelated Caucasian subjects. For the neuroimaging data, the reconstructed and attenuation-corrected DaT-SPECT images were aligned onto the standard MNI space and used to compute the specific binding ratio (SBR). The SBR is computed as the ratio between the concentrations of the specific binding radioactivity to the nonspecific binding radioactivity. The SBR is computed as the ratio between the target region and the reference region. We used the occipital cortex as the reference region. The SBR map of each subject was computed for all voxels, which were averaged to 90 region-of-interest (ROI) measurements using ROIs defined in the automated anatomical labeling (AAL) atlas.

We performed quality control of the genotyping data based on the minor allele frequency, genotype missing rate, Hardy–Weinberg equilibrium, and genotyping rate. SNPs that did not satisfy the criteria and subjects with a low genotyping rate were excluded. We then conducted the conventional genome-wide association analysis to select the candidate PD-related genes. Quality control analysis of the genetic data led to 148,631 SNPs. Conventional analysis to select candidate PD-related SNPs (corrected p-value < 0.001) led to 3,365 SNPs.

3. RESULTS

We quantified the performance of our algorithm on how well we can detect imaging genetics features (SNPs and regional DaT-SPECT). We compared our approach with the existing SCCA method. We first compared the variable selection performance by using simulation data. Our algorithm was applied to real PD imaging genetics data to investigate the brain regions potentially linked with PD and discover genetic variants associated with PD.

3.1. Simulation setup and results

3.1.1. Simulation setup

We used simulation datasets to compare the performance of DN-FS model to the existing SCCA algorithm. Specifically, we generated two correlated datasets with known ground truth of selection vectors (**Fig. 2-a**). A dataset of SNP data $\boldsymbol{X}$ with p SNPs and neuroimaging data $\boldsymbol{Y}$ (DaT-SPECT) with q regional features for n samples were generated. We applied a latent model to generate correlations between SNP and neuroimaging data [7], [8]. The correlations between SNPs and imaging features were generated with a latent variable ζ with a normal distribution $N(0, \sigma_\zeta^2)$. We generated one genetic selection vector $\boldsymbol{\alpha}$ with p elements and neuroimaging selection vectors $\boldsymbol{\beta}$ with q elements. We assumed that if an element of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ was correlated, the element was a non-zero value obtained from a uniform distribution U(-1, -0.5) ∪ U(0.5, 1). Otherwise, the elements would be zero. Each selection vector was made block sparse to mimic real data.

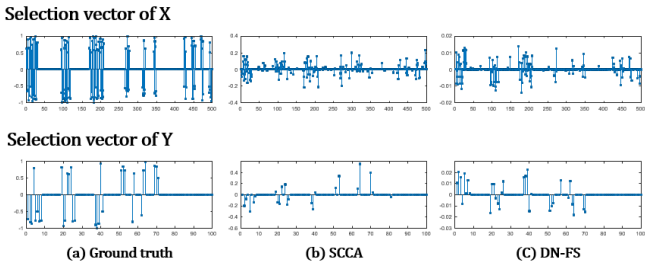

Fig. 2. Comparison of estimated selection vectors using simulation datasets. Each subfigure has two rows, where the first row is the selection vector of $\boldsymbol{X}$ (i.e., SNPs) and the second row is the selection vector for $\boldsymbol{Y}$ (i.e., DaT-SPECT). Subfigures (a), (b), and (c) correspond to the ground truth vectors and estimated selection vectors using the SCCA and DN-FS, respectively.

For neuroimaging data, we generated simulation datasets using $\boldsymbol{Y} = \boldsymbol{\beta\zeta} + \boldsymbol{e}$ for correlated features and $\boldsymbol{Y} = \boldsymbol{e}$ for uncorrelated ones, where noise e is drawn from the normal distribution $N(0, \sigma_e^2)$ with σ_e^2 as the noise variance. For genotyping data, we generated SNP variables by $\boldsymbol{X} = \boldsymbol{\alpha\zeta} + \boldsymbol{e}$ for correlated SNPs and $\boldsymbol{X} = \boldsymbol{e}$ for uncorrelated SNPs. Since the SNP data were categorical variables at three levels (0 [no minor allele], 1 [one minor allele], and 2 [two minor alleles]), we convert the SNP data $\boldsymbol{X}$ into categorical variables using a binomial distribution $\mathrm{B}(2, \mathrm{logit}^{-1}(\boldsymbol{X} + \mathrm{logit}(\eta)))$, where $\mathrm{logit}(\eta) = \log\left(\frac{\eta}{1-\eta}\right)$ was the logit function and the minor allele frequency η was drawn from a uniform distribution U(0.2, 0.4).

2.3.2. Simulation results

We compared the performance of DN-FS model with the existing SCCA method. The model performances of the algorithms were evaluated using the area under the curve (AUC) from the test data. We generated training and test data, where we set n = 200, p = 500, and q = 100 with a noise level of 3 (σ_e = 3), and n = 100, p = 500, and q = 100 with a noise level of 3 (σ_e = 3), respectively. Our algorithm showed improved performance for estimating the selection vectors compared to the SCCA method (**Fig. 2**). In detail, the AUC of our model showed 11% improvement over the SCCA (i.e., 0.8475 for our model and 0.7598 for SCCA). Both algorithms showed generally good performance for estimating genetic selection vectors (i.e., 0.8283 for our model and 0.8247 for SCCA), and our algorithm was better at estimating neuroimaging selection vectors (i.e., 0.9577 for our model and 0.7537 for SCCA).

3.2. Experiment on a real imaging genetics dataset

In addition to the experiment with simulation dataset, we tested the algorithms using a real dataset. We applied the algorithms to real genotyping and neuroimaging datasets of PD obtained from the PPMI research database. The

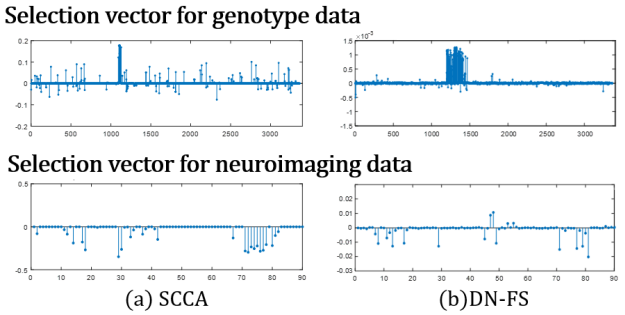

Fig. 3. Comparison of estimated selection vectors using real imaging genetics data. Each subfigure has two rows, where the first row is the selection vector for the SNPs and the second row is the selection vector for the neuroimaging data. Subfigures (a) and (b) correspond to the estimated selection vectors using the SCCA and DN-FS, respectively.

genotyping and neuroimaging data of 94 subjects with PD were used. We extracted 3,365 SNP markers and 90 regional DaT-SPECT features by using the preprocessing steps as described in Section 2.3.

We applied our algorithm to the preprocessed feature data (i.e., SNP and DaT-SPECT features). Our algorithm revealed that 181 SNPs were significantly linked with 22 ROIs of DaT-SPECT features and a canonical correlation of 0.6836 was reported between the identified genetic and DaT-SPECT features (**Fig. 3**). The SCCA algorithm showed that 197 SNPs and 29 DaT-SPECT features were significantly related and a canonical correlation of 0.6025 was reported between the identified features.

We further analyzed our results to determine whether the identified SNPs were consistent with the prior knowledge of PD. We compared the identified SNPs to PD-related SNPs in the PDGene database [9]. Using our algorithm, there were four genes, SNCA, IRF4, GCH1, and HCA-DQA1, which confirmed the existing findings in PD literature. The SNCA is one of the critical risk genes for PD. The SNCA gene provides instructions for making an alpha-synuclein, which plays an important role in movement structures. Previous reports showed that the mutations in SNCA were responsible for autosomal dominant PD [10]. In addition, we computed odd ratios (ORs) for the identified SNPs to measure the association between the identified SNPs and clinical diagnosis. We found that the identified SNPs by our algorithm showed an averaged OR of 1.53, which was higher than the values obtained using the competing method (0.96 for the SCCA). For the neuroimaging features, we found that the identified five ROIs were common in both methods, including globus pallidus, thalamus, Heschel's gyrus, bilateral inferior frontal gyrus, and pars opercularis. Most of these regions were parts of basal ganglia structures, which are related to PD [11].

4. CONCLUSION

In this study, we proposed a deep network-based feature selection model for evaluating associations between genetic and neuroimaging data. We demonstrated the feasibility of the algorithm using both simulation and real data. The proposed algorithm is an unsupervised learning model with deep learning framework. Thus, it could be used in unlabeled or categorized datasets. Our algorithm confirmed not only several well-established PD biomarkers but also revealed new potential SNPs. The proposed imaging genetics model has the potential to reveal new biomarkers to improve mechanistic understanding of the studied brain disorders. In the future, we will expand our method to the analysis of additional modalities, and apply our method to cohorts such as Alzheimer's patients and the healthy to test if our approach generalizes to other cases.

5. REFERENCES

[1] X. Zhu, H.-I. Suk, H. Huang, and D. Shen, "Low-Rank Graph-Regularized Structured Sparse Regression for Identifying Genetic Biomarkers," *IEEE Trans. Big Data*, vol. 7790, no. c, pp. 1–1, 2017.

[2] X. Yao *et al.*, "Tissue-specific network-based genome wide study of amygdala imaging phenotypes to identify functional interaction modules," *Bioinformatics*, vol. 33, no. 20, pp. 3250–3257, 2017.

[3] H. Wang *et al.*, "Identifying quantitative trait loci via group-sparse multitask regression and feature selection: An imaging genetics study of the ADNI cohort," *Bioinformatics*, vol. 28, no. 2, pp. 229–237, 2012.

[4] Y. Li, C. Y. Chen, and W. W. Wasserman, "Deep feature selection: Theory and application to identify enhancers and promoters," *J. Comput. Biol.*, vol. 23, no. 5, pp. 322–336, 2016.

[5] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.

[6] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017–Octob, pp. 2242–2251, 2017.

[7] J. Fang, D. Lin, S. C. Schulz, Z. Xu, V. D. Calhoun, and Y. P. Wang, "Joint sparse canonical correlation analysis for detecting differential imaging genetics modules," *Bioinformatics*, vol. 32, no. 22, pp. 3480–3488, 2016.

[8] D. Lin, V. D. Calhoun, and Y. P. Wang, "Correspondence between fMRI and SNP data by group sparse canonical correlation analysis," *Med. Image Anal.*, vol. 18, no. 6, pp. 891–902, 2014.

[9] C. M. Lill *et al.*, "Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: The PDGene database," *PLoS Genet.*, vol. 8, no. 3, p. e1002548, 2012.

[10] I. J. Siddiqui, N. Pervaiz, and A. A. Abbasi, "The Parkinson Disease gene SNCA: Evolutionary and structural insights with pathological implication," *Sci. Rep.*, vol. 6, no. March, pp. 1–11, 2016.

[11] M. Kim, J. Kim, S. H. Lee, and H. Park, "Imaging genetics approach to Parkinson's disease and its correlation with clinical score," *Sci. Rep.*, vol. 7, p. 46700, 2017.