2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Use of deep learning genomics to discriminate Alzheimer's disease and healthy controls

Lanlan Li, Yanru Huang, Ying Han, Jiehui Jiang*, *Member, IEEE,* and the Alzheimer's Disease Neuroimaging Initiative

***Abstract*—Alzheimer's disease (AD) is the most prevalent neurodegenerative disorder and the most common form of dementia in the elderly. Because gene is an important clinical risk factor resulting in AD, genomic studies, such as genome-wide association studies (GWAS), have widely been applied into AD studies. However, main shortcomings of GWAS method were that hereditary deletions were evident in the GWAS studies, which resulted in low classification or prediction abilities by using GWAS analysis. Therefore, this paper proposed a novel deep learning genomics approach and applied it to discriminate AD patients and healthy control (HC) subjects. In this study, we selected genotype data of 988 subjects enrolled in the ADNI, including 622 AD patients and 366 HC subjects. The proposed deep learning genomics (DLG) approach was composed of three steps: quality control, SNP genotype coding, and classification. The Resnet framework was used as the DLG model in this study. In the comparative GWAS analysis, APOE ε4 status and the normalized theta-value of the significant SNP loci were seen as predictors to classify genetically using Support Vector Machine (SVM). All data were divided into one training & validation group and one test group. 5-fold cross-validation was used in 500 times. Finally, we compared the classification results between DLG model and traditional GWAS analysis. As a result, the accuracy, sensitivity, and specificity of classification for traditional GWAS analysis was 71.38%±0.63%, 63.13%±2.87% and 85.59%±6.66% in the test group; while the accuracy, sensitivity, and specificity of classification for DLG model was 92.65%±4.80%, 85.00%±16.25% and 97.10%±4.38% in the test group. Hence, the DLG model can achieve higher accuracy and sensitivity when applied to AD. More importantly, we discovered several novel genetic biomarkers of AD, including rs6311 and rs6313 in HTR2A, and rs690705 in RFC3. The roles of these novel loci in AD should be explored future.**

***Keywords*—Alzheimer's disease, deep learning genomics, genome-wide association studies (GWAS), computed aided diagnosis**

I. Introduction

Alzheimer's disease (AD) is the most common type of dementia and is an irreversible, progressive brain disorder typically beginning with mild memory loss; later it can seriously impair an individual's ability to carry out daily activities. It has been widely recognized and emphasized that early detection of AD is beneficial.

Among factors that influence AD progression, common genetic variants are the major risk factors [1]. Right now, sequencing and omics analysis techniques have been widely used in this issue, such as advanced genome-wide association studies (GWAS) and whole genome sequencing (WGS) studies [2]. For instance, APOE was proven as the most strongly associated AD risk gene in the omics analysis. Besides, recent studies of the Alzheimer's Disease Neuroimaging Initiative (ADNI) GWAS data have related known AD risk genes to differences in rates of brain atrophy and biomarkers of AD in the cerebrospinal fluid [3]. Therefore, omics analysis, especially GWAS analysis, has shown general advances in AD research.

However, it is still under exploration on how to analyze AD progression utilizing original genomics data. First of all, most GWAS studies focused to identify significant genetic loci and used these significant loci for future analysis. However, these loci that showed strongest associations with the disease may be not generally the causal Single Nucleotide Polymorphisms (SNP). Secondly, hereditary deletions were evident in the GWAS studies, which resulted in low classification or prediction abilities of the disease by using GWAS analysis. Thirdly, traditional GWAS analysis required a plenty of prior knowledge and hand coding, which results in relatively low distinction and the exhaustion of time and energy. Therefore, alternative analytical tools were required to drive novel hypotheses and models in this topic.

Deep learning algorithms can embed the computation of features automatically to yield end-to-end models to discover relevant features of high complexity [4]. In recent studies, deep convolutional neural networks have been used to predict various molecular phenotypes on the basis of DNA sequence alone,such as classifying transcription factor binding sites, predicting molecular phenotypes such as DNA methylation and gene expression [5,6]. Hence, in this work, we aim to propose a deep learning genomics (DLG) approach to replace traditional GWAS analysis and apply DLG to seek novel genetic biomarkers of AD susceptibility.

II. Materials and Methods

A. Experimental framework

The workflow of this study was shown in Fig. 1, which was composed of three steps. First, we processed quality control and conducted SNP genotype coding for satisfactory SNP genotype data. Second, we presented the deep residual network Resnet34 for the transfer learning of DLG. The goal of the deep residual network was to obtain a model by supervised training for prediction and extract DLG features.

*Research supported by the National Natural Science Foundation of China No.61603236, 61633018, Shanghai Municipal Science and TechnologyMajor Project (Grant No. 2017SHZDZX01).

Lanlan Li, Yanru Huang, Jiehui Jiang are with the Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China (corresponding author to provide phone: +86 139-1892-0926; e-mail: jiangjiehui@shu.edu.cn).

Ying Han is with Center of Alzheimer's Disease, Institute for Brain Disorders, Beijing, China.

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

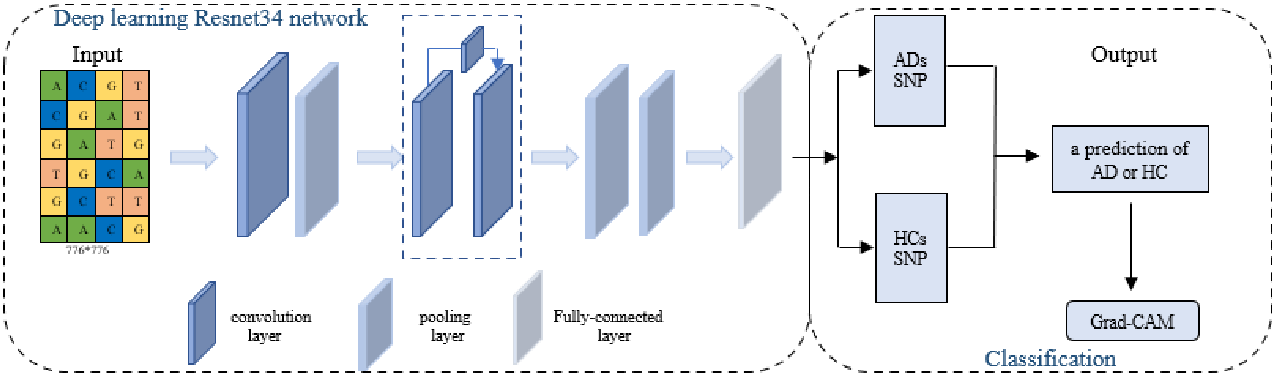

Figure 1. The framework of this study.

B. Materials

Data used in the preparation of this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/ADNI). In this study, 988 (AD = 622, healthy controls =366) individuals from the ADNI cohort were included. Meanwhile, the following data from 988 ADNI participants were downloaded from ADNI, including the Illumina SNP genotyping data, demographic information and diagnosis information. Clinical characteristics including age, sex, education and Mini-Mental State Examination (MMSE) were listed in Table 1.

TABLE I. DEMOGRAPHIC AND CLINICAL CHARACTERISTICS

	HC (n = 366)	*AD (n = 622)*
Female/Male	189/177	261/361
Age	73.96±5.67	74.55±7.33
Education	16.38±2.67	15.53±2.90[a]
MMSE score	29.07±1.12	22.31±4.03[a]

Note: Age, Education, MMSE are given as mean ± standard deviation.

a Two sample *t*-test, p < 0.05, HC and AD.

C. DNA isolation and SNP genotyping

SNP genotyping for more than 620000 target SNPs was completed on all ADNI participants using the following protocol. First, a total of 7 mL of blood was taken in EDTA-containing Vacutainer tubes from all participants and genomic DNA was extracted using the QIAamp DNA Blood Maxi Kit following the manufacturer's protocol. Second, lymphoblastoid cell lines were established by transforming B lymphocytes with Epstein-Barr virus. 14 Genomic DNA samples were analyzed using the Human 610-Quad BeadChip according to the manufacturer's protocols. Before initiation of the assay, 50 ng of genomic DNA from each sample was examined qualitatively on a 1% Tris-acetate-EDTA agarose gel to check for degradation. Degraded DNA samples were excluded from further analysis. Third, samples were quantitated in triplicate with PicoGreen® reagent and diluted to 50 ng/L in TrisEDTA buffer (10 mM Tris, 1 mM EDTA, pH 8.0). A total of 200 ng of DNA was then denatured, neutralized, and amplified for 22 hours at 37°C (this is termed the MSA1 plate). The MSA1 plate was fragmented with FMS reagent (Illumina) at 37°C for 1 hour, precipitated with 2-propanol, and incubated at 4°C for 30 minutes. Fourth, the resulting blue precipitate was re-suspended in RA1 reagent (Illumina) at 48°C for 1 hour. Samples were then denatured (95°C for 20 minutes) and immediately hybridized onto the BeadChips at 48°C for 20 hours. The BeadChips were washed and subjected to single base extension and staining. Finally, the BeadChips were coated with XC4 reagent (Illumina), desiccated, and imaged on the BeadArray Reader (Illumina). The Illumina BeadStudio 3.2 software was used to generate SNP genotypes from bead intensity data [7].

D. Quality control and APOE genotype

The following quality control (QC) steps were performed on these 998 samples using the PLINK software package (http://pngu.mgh.harvard.edu/~purcell/plink/), release v1.07. SNPs and participants were excluded from the analysis if they could not meet any of the following criteria [8]: Call rate per SNP ≥90%; Call rate per participant ≥90%; Gender check; Minor allele frequency (MAF) ≥5%; Hardy–Weinberg equilibrium test of $p \leq 10^{-6}$; PI_HAT<0.5. After the QC procedure, 301388 features for each subject were considered for further analysis. The overall genotyping rate for the remaining dataset was over 99.5%.

Although the APOE gene was an important target gene in AD research, it was not available for all identified APOE SNPs on the Illumina array. Therefore, the genotypes of the APOE SNPs that were not available were added to ADNI genotype data based on the reported the APOE ε2/ε3/ε4 status before the assessment of sample quality.

E. SNP genotype coding

A single nucleotide polymorphism is a DNA sequence variation occurring when a single nucleotide (A, T, C, or G) in the genome differs among members of a biological species or across paired chromosomes. Based on the ADNI GWAS SNP data, in this study, we encoded SNPs using the coding scheme as follows: A refers to 1, T refers to 2, C refers to 3, and G refers to 4.

F. GWAS analysis

GWAS has been emerged as a popular tool to identify genetic variants that are associated with disease risk. Standard analysis of a case-control GWAS involves assessing the association between each individual genotyped SNP and disease risk. A Manhattan plot and a quantile–quantile (Q–Q) plot were used to visualize GWAS results. All association results surviving the significance threshold of $p < 1.66e - 7$ were saved and prepared for additional pattern analysis.

G. DLG model

The based DLG model acted as a feature encoder, which had a significant impact on classification. In this study, we applied Resnet34 model to the classification between AD and

HC groups. The greatest advantage of Resnet framework lies in adding identity mapping that is performed by the shortcut connections and their outputs are added to the outputs of the stacked layers. Therefore, the Resnet addressed the degradation problem and added neither extra parameter nor computational complexity. The formula of residual learning was designed as: denoting the desired underlying mapping as $H(x)$, we let the stacked nonlinear layers fit another mapping of $F(x) = H(x) - x$. The original mapping was recast into $F(x) + x$. The formulation of $F(x) + x$ can be realized by feedforward neural networks with "shortcut connections" [9]. Fig. 2 showed the building block of the residual learning model.

There were two steps included in the entire process, the forward computation and the backward propagation. Before that, each subject's SNP genotype data were cropped after quality control, generated to 776 ×776 pixels. In the training stage, SNP genotype data were fed into the network to update model parameters by backward propagation. The outputs of the network were used as the classification results, and the cross-entropy of the outputs were calculated as the loss function. We set learning rate to 1e-3 and applied the Adam optimizer to update the model parameters with batch size 10. The maximum iteration step was set to 20.

For investigating the interpretability of the DLG model, the last convolutional layer of the last res-block was made transparent to extract DLG features by applying the Gradient-weighted Class Activation Mapping (Grad-CAM).

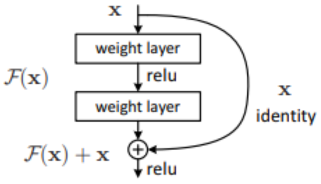
x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2. Residual learning: a building block.

H. Classification

The enrolled subjects were randomly divided into one training group and one independent test group with the ratio of 9:1. The training group was then used to optimize the model parameters. We also randomly chose 25% of training group to form a validation group to guide the choice of hyper parameters.

To verify the diagnostic capabilities of the DLG model compared with traditional GWAS analysis, we performed comparative trials. Among all the gene indicators, theta value was proved to have the most direct relationship with SNP changes. APOE ε4 status and the normalized theta-value of the significant SNP loci found in this study were seen as predictors to classify genetically and we used Support Vector Machine (SVM) with the linear kernel 500 times for classification.

To evaluate classification performance, we repeatedly conducted 5-fold cross-validation in the training group and verified in the test group. Accuracy, sensitivity, and specificity of the test group were used to evaluate the results.

I. Statistical analysis

Demographic characteristics were compared based on two-sample t test or the chi-square test. Two-sample t test among features extracted was applied as a criterion to estimate the differences of DLG features between AD patients and HCs. All statistical analyses were performed in SPSS Version 22.0 software (SPSS Inc., Chicago, IL). All p value < 0.05 was considered significant.

III. Results

A. Outcomes of GWAS analysis

After GWAS analysis, we observed two genome-wide significant loci on chromosome 19, including rs429358 (APOE, the epsilon 4 marker) and rs2075650 (TOMM40). Fig. 3 showed the Manhattan and Q–Q plots of the GWAS analysis.

B. Classification performance

Table 2 showed the classification accuracy, sensitivity, specificity and area under curve (AUC) of the GWAS analysis and the DLG model. In the test group, the GWAS analysis could achieve accuracy, sensitivity, specificity and AUC of 71.38%±0.63%, 63.13%±2.87%, 85.59%±6.66% and 0.744. The DLG model achieved the accuracy, sensitivity, specificity and AUC of 92.65%±4.80%, 85.00%±16.25%, 97.10%±4.38% and 0.999. As a result, the DLG model was more superior to the traditional GWAS analysis for classification.

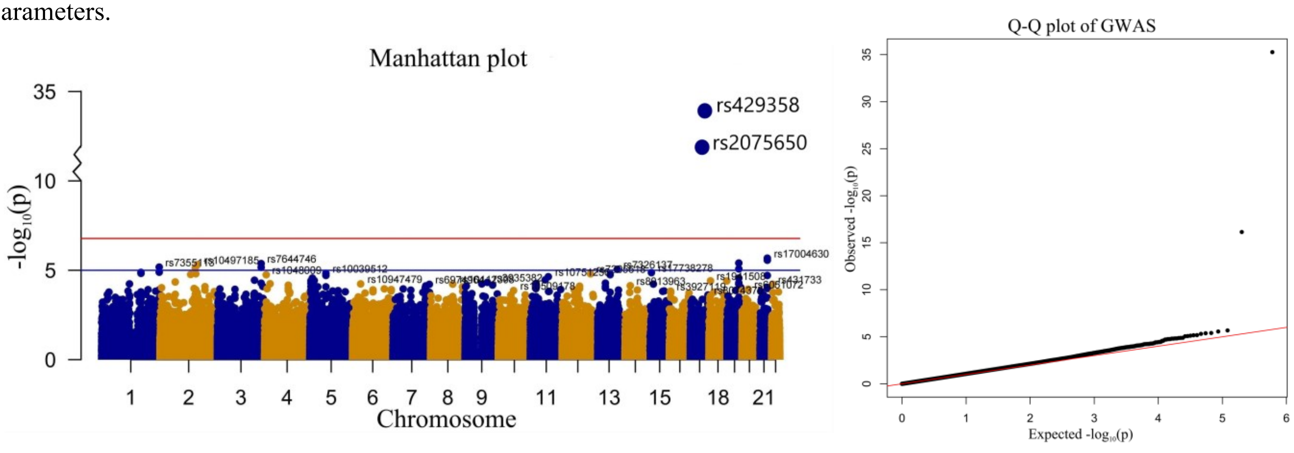

Figure 3. Manhattan and Q–Q plots of genome-wide association study (GWAS). The horizontal lines in the Manhattan plot display the cutoffs for two significant levels: blue line for $p < 10^{-5}$, and red line for $p < 1.66\mathrm{e} - 7$. Genomic inflation factor is 1.084.

TABLE II. PERFORMANCE OF DIFFERENCE CLASSIFICATION APPROACHES

	Accuracy (%)	*Sensitivity (%)*	*Specificity (%)*	*AUC*
GWAS analysis	71.38±0.63	63.13±2.87	85.59±6.66	0.744
DLG model	92.65±4.80	85.00±16.25	97.10±4.38	0.999

Note: The methods are conducted with cross-validation and their results are given as mean ±standard deviation.

C. Interpretability of the DLG model

The interpretability of DLG model was explored based on the Grad-CAM and two-sample t test. Setting the threshold for p value <0.05, feature information of more over ten thousand SNP loci showed differences between AD and HC groups. Table 3 showed several SNP loci that were found through the DLG model.

TABLE III. RESULTS OF THE DLG MODEL INTERPRETABILITY

SNP	***CH***	***Region or Closest Gene***	***P value***
rs16847609	3	SOX14/CLDN18	0.039
rs2067477	11	CHRM1;LOC105369333	0.013
rs690705	13	RFC3	0.045
rs6311	13	HTR2A	0.021
rs6313	13	HTR2A	0.027
rs2073475	15	CYP11A1	0.010
rs2456930	15	TLN2	0.010

Note: SNP, singlenucleotide polymorphism. CH = chromosome.

IV. DISCUSSION

This paper proposed a deep learning genomics approach based on Resnet34. The classification experiment results indicated the higher diagnosis value of the DLG model compared with traditional GWAS analysis.

In GWAS analysis two SNPs were identified at the $p < 1.66e - 7$ significance level. As a well-established AD risk factor, the APOE SNP rs429358 was determined as the most prominent genetics. Moreover, the second significant TOMM40 SNP rs2075650 was also found as a gene adjacent to APOE and an additional contributor to AD [10]. These results were consistent with previous studies [3,10].

Besides, when we interpreted the DLG model, we found more over one thousand SNP loci with significant difference. Among those, the A allele of rs16847609 had been reported to be associated with AD in APOE ε4- carriers [11]. In addition, rs2456930 was revealed to influence temporal lobe structure with relevance to neurodegeneration in Alzheimer's disease [12]. The SNP loci rs690705 was also a characteristic of important GWAS SNPs associated with AD [13]. That was to say, the DLG model indeed had the ability to identify the difference of genomics between AD and HC groups.

It was worth noting that this study had some limitations. Firstly, only gene sequences were used as the inputs of DLG for classification. We planned to combine gene sequences with clinical data and brain imaging together to facilitate the classification abilities of DLG. Secondly, the materials in this study were limited. We only compared the classification results between AD and HC groups in this study. We would like to test our model in other datasets such as mild cognitive impairment in the future. Thirdly, we only deployed one kind of DLG models in this study. We would like to utilize different deep learning models and compare the classification results to choose the best. Lastly, the dataset used in this study may not be large enough. The results of this study need to be further verified by other datasets.

In conclusion, this study suggested that the DLG approach was effective in AD research and outperformed traditional GWAS analysis. Moreover, the several novel SNP loci identified in the DLG approach including rs6311 and rs6313 in HTR2A, and rs690705 in RFC3 will be worthy of further exploration to better understand the mechanisms of AD.

REFERENCES

[1] Y. Huang and L. Mucke, “Alzheimer mechanisms and therapeutic strategies.” Cell, vol. 148, no. 6, pp. 1204-1222, 2012.

[2] C. Sarnowski et al., “Whole genome sequence analyses of brain imaging measures in the Framingham Study.” Neurology, vol. 90, no. 3, pp. e188-e196, 2018.

[3] S. Kim et al., “Genome-wide association study of CSF biomarkers Abeta1-42, t-tau, and p-tau181p in the ADNI cohort.” Neurology, vol. 76, no. 1, pp. 69-79, 2011.

[4] G. Eraslan, Z. Avsec, J. Gagneur, and F. J. Theis, “Deep learning: new computational modelling techniques for genomics.” Nat Rev Genet, vol. 20, no. 7, pp. 389-403, 2019.

[5] H. Zeng and D. K. Gifford, “Predicting the impact of non-coding variants on DNA methylation,” Nucleic acids research, vol. 45, no. 11, pp. e99, 2017.

[6] J. Zhou et al., “Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk,” Nat Genet, vol. 50, no. 8, pp. 1171-1179, 2018.

[7] H. Neitzel, “A routine method for the establishment of permanent growing lymphoblastoid cell lines,” Human Genetics, vol. 73, no. 4, pp. 320–326, 1986.

[8] L. Shen et al., “Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort.” NeuroImage, vol. 53, no. 3, pp. 1051-1063, 2010.

[9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition.” IEEE computer society, pp. 770-778, 2016.

[10] H. Huang et al., “The TOMM40 gene rs2075650 polymorphism contributes to Alzheimer's disease in Caucasian, and Asian populations.” Neurosci Lett, vol. 628, pp. 142-146, 2016.

[11] G. Jun et al., “A novel Alzheimer disease locus located near the gene encoding tau protein.” Mol Psychiatry, vol. 21, no. 1, pp. 108-117, 2016.

[12] J. L. Stein et al., “Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in Alzheimer's disease.” NeuroImage, vol. 51, no. 2, pp. 542-554, 2010.

[13] C. F. Moraes, T. C. Lins, E. F. Carmargos, J. O. Naves, R. W. Pereira, and O. T. Nobrega, “Lessons from genome-wide association studies findings in Alzheimer's disease.” Psychogeriatrics, vol. 12, no. 1, pp. 62-73, 2012.