Open access • Posted Content • DOI:10.1101/2021.07.19.21260789

Deep learning-based identification of genetic variants: Application to Alzheimer's disease classification — Source link

Taeho Jo, Kwangsik Nho, Paula J. Bice, Andrew J. Saykin ...+1 more authors

Institutions: Indiana University

Published on: 22 Jul 2021 - medRxiv (Cold Spring Harbor Laboratory Press)

Topics: Deep learning

Deep learning-based identification of genetic variants: Application to Alzheimer's disease classification

Taeho Jo[1,2,3], Kwangsik Nho[1,2,3,4*], Paula Bice[1,2], and Andrew J. Saykin[1,2,3,5*], for the Alzheimer's Neuroimaging Initiative [¥]

[1]Department of Radiology and Imaging Sciences, Center for Neuroimaging, Indiana University School of Medicine, Indianapolis, IN, USA

[2]Indiana Alzheimer's Disease Research Center, Indiana University School of Medicine, Indianapolis, IN, USA

[3]Indiana University Network Science Institute, Bloomington, IN, USA

[4]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, USA

[5]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA

*Corresponding authors

Kwangsik Nho, PhD

Department of Radiology and Imaging Sciences and the Indiana Alzheimer's Disease Research Center,
Indiana University School of Medicine,
355 W 16th St. IU Neuroscience Center, GH 4101, Indianapolis, Indiana 46202, USA
Tel: 317-963-7503; Fax: 317-274-1067; E-mail: knho@iupui.edu

Andrew J. Saykin, PsyD

Department of Radiology and Imaging Sciences and the Indiana Alzheimer's Disease Research Center,
Indiana University School of Medicine, Indianapolis, IN, USA,
355 W 16th St. IU Neuroscience Center, GH 4101, Indianapolis, Indiana 46202, USA
Phone: 317-278-6947; Fax: 317-274-1067; E-mail: asaykin@iupui.edu

[¥]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Email addresses:

Taeho Jo: tjo@iu.edu

Kwangsik Nho: knho@iupui.edu

Paula Bice: pbice@indiana.edu

Andrew J Saykin: asaykin@iupui.edu

Abstract

Deep learning is a promising tool that uses nonlinear transformations to extract features from high-dimensional data. Although deep learning has been used in several genetic studies, it is challenging in genome–wide association studies (GWAS) with high-dimensional genomic data. Here we propose a novel three-step approach for identification of genetic variants using deep learning to identify phenotype-related single nucleotide polymorphisms (SNPs) and develop accurate classification models. In the first step, we divided the whole genome into non-overlapping fragments of an optimal size and then ran Convolutional Neural Network (CNN) on each fragment to select phenotype-associated fragments. In the second step, using an overlapping window approach, we ran CNN on the selected fragments to calculate phenotype influence scores (PIS) and identify phenotype-associated SNPs based on PIS. In the third step, we ran CNN on all identified SNPs to develop a classification model. We tested our approach using genome-wide genotyping data for Alzheimer's disease (AD) (N=981; cognitively normal older adults (CN) =650 and AD=331). Our approach identified the well-known *APOE* region as the most significant genetic locus for AD. Our classification model achieved an area under the curve (AUC) of 0.82, which outperformed traditional machine learning approaches, Random Forest and XGBoost. By using a novel deep learning-based GWAS approach, we were able to identify AD-associated SNPs and develop a better classification model for AD.

Author summary

Although deep learning has been successfully applied to many scientific fields, deep learning has not been used in genome–wide association studies (GWAS) in practice due to the high dimensionality of genomic data. To overcome this challenge, we propose a novel three-step approach for identification of genetic variants using deep learning to identify disease-associated single nucleotide polymorphisms (SNPs) and develop accurate classification models. To accomplish this, we divided the whole genome into non-overlapping fragments of an optimal size and ran a deep learning algorithm on each fragment to select disease-associated fragments. We calculated phenotype influence scores (PIS) of each SNP within selected fragments to identify disease-associated significant SNPs and developed a disease classification model by using overlapping window and deep learning algorithms. In the application of our method to Alzheimer's disease (AD), we identified well-known significant genetic loci for AD and achieved higher classification accuracies than traditional machine learning methods. This study is the first study to our knowledge to develop a deep learning-based identification of genetic variants using fragmentation and window approaches as well as deep learning algorithms to identify disease-related SNPs and develop accurate classification models.

Introduction

Deep learning is a representative machine learning algorithm that enables nonlinear transformations to extract features of high-dimensional data [1], unlike traditional machine learning models that predict a linear combination of weights by assuming a linear relationship between input features and a phenotype of interest. Deep learning has been used to predict disease outputs by handling original high-dimensional medical imaging data without feature selection procedures [2, 3]. In genetic research, deep learning frameworks have been used to investigate molecular phenotypes that predict the effects of non-coding variants[4-10], differential gene expression [11], and potential transcription factor binding sites [12]. These tools use CHIP-Seq or DNase-Seq data as training data to predict chromatin features such as transcription factor binding or DNase hypersensitivity from DNA sequences. More recently, deep learning has been employed in the capture of mutations and the analysis of gene regulations, demonstrating its potential for furthering our understanding of epigenetic regulation [13]. Furthermore, deep learning is being used in gene therapy to design CRISPR guide RNAs using deep learning-based gene features [14-19].

Genome-wide association studies (GWAS) use a statistical approach by considering one single nucleotide polymorphism (SNP) at a time across the whole genome to identify population-based genetic risk variation for human diseases and traits [20, 21]. However, deep learning has not yet been used to perform GWAS, as it is challenging due to the so-called high-dimension low-sample-size (HDLSS) problem [22], which is known to impact phenotype prediction using genetic variation. Feature reduction approaches have been commonly used [23-25] to resolve this problem, but feature reduction using high-dimensional genomic data is also challenging due to a NP-hard problem[26, 27]. Therefore, it is necessary to develop a deep learning framework to identify genetic variants using whole genome data.

Here we proposed a novel three-step deep learning-based approach to select informative SNPs and develop classification models for a phenotype of interest. In the first step, we divided the whole genome into non-overlapping fragments of an optimal size and then used deep learning algorithms to select phenotype-associated fragments containing phenotype-related SNPs. Different sized fragments and several deep learning algorithms were tested to select the optimal size for fragments and the optimal algorithm. In the second step, we ran the optimal deep learning algorithm using an overlapping sliding window approach within selected fragments to calculate phenotype influence scores (PIS) using SNPs and the phenotype of interest to identify informative SNPs. In the third step, we ran the optimal algorithm again on all identified informative SNPs to develop a classification model.

We tested our approach using only whole genome data for Alzheimer's disease (AD) (N=981; cognitively normal older adults (CN) =650 and AD=331). Our approach identified the known *APOE* region as the most significant genetic locus for AD. Our classification model yielded 75.2% accuracy over traditional machine learning methods, being 3.8% and 9.6% higher than XG Boost and Random Forest, respectively. Our novel deep learning-based approach can identify informative SNPs and develop a classification model for AD by combining nearby SNPs and testing their aggregation.

Materials and Methods

Study participants

All individuals used in the analysis were participants of the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort [28, 29]. The ADNI initial phase (ADNI-1) was launched in 2003 to test whether serial magnetic resonance imaging (MRI), position emission tomography (PET), other biological markers, and clinical and neuropsychological assessment could be combined to measure the progression of Mild Cognitive Impairment (MCI) and early AD. ADNI-1 has been extended in subsequent phases (ADNI-GO, ADNI-2, and ADNI-3) for follow-up of existing participants and additional new enrollments. Demographic information, *APOE* and whole-genome genotyping data, and clinical information are publicly available from the ADNI data repository (www.loni.usc.edu/ADNI/). Informed consent was obtained for all subjects, and the study was approved by the relevant institutional review board at each data acquisition site.

Genotyping and imputation

ADNI participants were genotyped using several Illumina genotyping platforms including Illumina Human610-Quad BeadChip, Illumina HumanOmni Express BeadChip, and Illumina HumanOmni 2.5M BeadChip [30]. As ADNI used different genotyping platforms, we performed quality control procedures (QC) on each genotyping platform data separately and then imputed un-genotyped single nucleotide polymorphisms (SNPs) separately using MACH and the Haplotype Reference Consortium (HRC) data as a reference panel [31]. Before imputation, we performed QC for samples and SNPs as described previously: (1) for SNP, SNP call rate < 95%, Hardy-Weinberg *P* value < 1×10^{-6}, and minor allele frequency (MAF) < 1%; (2) for sample, sex inconsistencies, and sample call rate < 95% [32]. Furthermore, in order to prevent spurious association due to population stratification, we selected only non-Hispanic participants of European ancestry that clustered with HapMap CEU (Utah residents with Northern and Western European ancestry from the CEPH collection) or TSI (Toscani in Italia) populations using multidimensional scaling (MDS) analysis and the HapMap genotype data [32, 33]. After imputation, we performed standard QC on imputed genotype data as described previously [34]. Specifically, we imposed an r^2 value equal to 0.30 as the threshold to accept the imputed genotypes. In the study, imputed genome-wide genotyping data from 981 ADNI non-Hispanic participants (650 cognitive normal older adults (CN) and 331 AD patients) were used with a total of 5,398,183 SNPs (minor allele frequency (MAF) > 5%).

Genome-wide association study (GWAS)

Using imputed genotypes, a GWAS for AD was conducted. For the GWAS, logistic regression with age and sex as covariates was performed using PLINK[35] to determine the association of each SNP with AD. To adjust for multiple testing, a conservative threshold for genome-wide significant association ($p < 5 \times 10^{-8}$) was employed based on a Bonferroni correction.

Fragmentation of whole genome data

Whole genome data for 981 participants were divided into non-overlapping fragments of varying sizes from 10 SNPs to 200 SNPs to determine the optimal fragmentation size. The sub-data sets consisting of fragments of the same size were divided into train-test-validation sets (60:20:20), and Convolutional Neural Network (CNN)[36], Long short-term memory (LSTM)[37], LSTM-CNN[38], and Attention[39] algorithms were applied to each. Early stopping using a validation set was applied to prevent over-fitting, followed by the measurement of training time and accuracy (ACC).

Deep learning on fragments

Deep learning is the result of continuous development such as perceptron[40, 41], which adds the concept of weight adjustment to the theory that it can behave like a human brain when neurons with on-off functions are connected in a network form[42], and Adaline[43], which uses gradient descent to update weights. These early neural nets were advanced to a multilayer perceptron, which includes hidden layers to solve the famous XOR problem[44], marking a theoretical turning point with the concept of backpropagation to update the weight of the hidden layer[45-48]. The inherent problem of backpropagation, in which vanishing gradients occur when there are many layers[49], has been alleviated through activation functions, such as sigmoid function and ReLU[50, 51], and optimization methods for better gradient descent methods, such as Ada-Grad[52], RMSprop[53], and Adam[54]. These developments, along with the advancement of GPU hardware, have created an era of deep learning as it is now.

Deep learning has laid the theoretical foundation for backpropagation, the application of activation functions, and the development of optimizers for better gradient descent. Common deep learning algorithms, such as CNN, LSTM, and Attention, have a hierarchical structure that implements an enhanced version of the basic principles of deep learning. The detailed technical description of each algorithm is described extensively in the relevant paper, so here we focus on the core of the deep learning technology commonly applied to the algorithm used in the experiment.

We used ReLU as an activation function that underlies the deep learning algorithms used in our experiments.

$$ReLU(x) = \begin{cases} x & if\ x \geq 0 \\ 0 & if\ x < 0 \end{cases}$$

ReLU, the most used activation function in the deep learning community, replaces the given value with zero if the value is < 0 and uses the given value if it is > 0. Thus, if the given value is greater than zero, the derivative becomes one, and the weight can be adjusted without vanishing the gradient to the first layer through the hidden layer. We used Adam as the optimization method. Adam, is currently the most popular optimization method for deep learning, as it takes advantage of momentum SGD[55] and RMSprop, which are expressed as follows: G_t is the sum of the square of the modified gradient, and ε is a very small constant that prevents the equation from being divided by zero.

$$\boldsymbol{V_t = \gamma G_{(t-1)} + (1 - \gamma_1) \frac{\partial Error}{\partial W_t}}$$

$$G_t = \gamma G_{(t-1)} + (1-\gamma_2)\left(\frac{\partial Error}{\partial W_t}\right)^2$$

$$\widehat{V}_t = \frac{V_t}{1-\gamma_1^t} \quad \widehat{G}_t = \frac{G_t}{1-\gamma_2^t}$$

$$W_{(t+1)} = W_t - \eta\frac{\widehat{G}_t}{\sqrt{\widehat{V}_t + \epsilon}}$$

Backpropagation is used to calculate the initial error value from a given random weight using the least squares method and then to update the weight using a chain rule until the differential value becomes zero. Here, the differential value of zero means that the weight does not change when the gradient is subtracted from the previous weight.

$$W_o(t+1) = W_o t - \frac{\partial ErrorY_o}{\partial W_o}$$

$$ErrorY_o = \frac{1}{2}(y_{t1} - y_{o1})^2 + \frac{1}{2}(y_{t2} - y_{o2})^2$$

If y_{o1} and y_{o2} are the output values of the output layer coming through the hidden layer, and the actual values of the given data are y_{t1} and y_{t2}, the partial derivative of the error $ErrorY_o$ to the weight of the output layer can be calculated using the chain rule as follows:

$$\frac{\partial ErrorY_o}{\partial w_o} = \frac{\partial ErrorY_o}{\partial y_{o1}} \cdot \frac{\partial y_{o1}}{\partial net3} \cdot \frac{\partial net3}{\partial w_o}$$

The partial derivative of the error $ErrorY_o$ to the weight of the hidden layer can be calculated as follows:

$$\frac{\partial ErrorY_o}{\partial h_1} = \frac{\partial (Errory_{o1} + Errory_{o2})}{\partial y_{h1}} = \underbrace{\frac{\partial Errory_{o1}}{\partial y_{h1}}}_{(a)} + \underbrace{\frac{\partial Errory_{o2}}{\partial y_{h1}}}_{(b)}$$

$$\text{(a)}\ \frac{\partial Errory_{o1}}{\partial y_{h1}} = \frac{\partial Errory_{o1}}{\partial net_3} \cdot \frac{\partial net_3}{\partial y_{h1}} = (y_{o1} - y_{t1})y_{o1}(1 - y_{o1})y_{o1}$$

$$\text{(b)}\ \frac{\partial Errory_{o2}}{\partial y_{h1}} = \frac{\partial Errory_{o2}}{\partial net_4} \cdot \frac{\partial net_4}{\partial y_{h1}} = (y_{o2} - y_{t2})y_{o2}(1 - y_{o2})y_{o2}$$

Accordingly, the weight w_h of the hidden layer is updated as follows:

$$\frac{\partial ErrorY_o}{\partial w_h} = \frac{\partial ErrorY_o}{\partial y_{h1}} \cdot \frac{\partial y_{h1}}{\partial net_1 y} \cdot \frac{\partial net_1}{\partial w_h}$$

$$= (\delta y_{o1} y_{o1} - \delta y_{o2} y_{o2}) y_{h1}(1 - y_{h1}) x_1$$

Calculation of phenotype influence score using deep learning

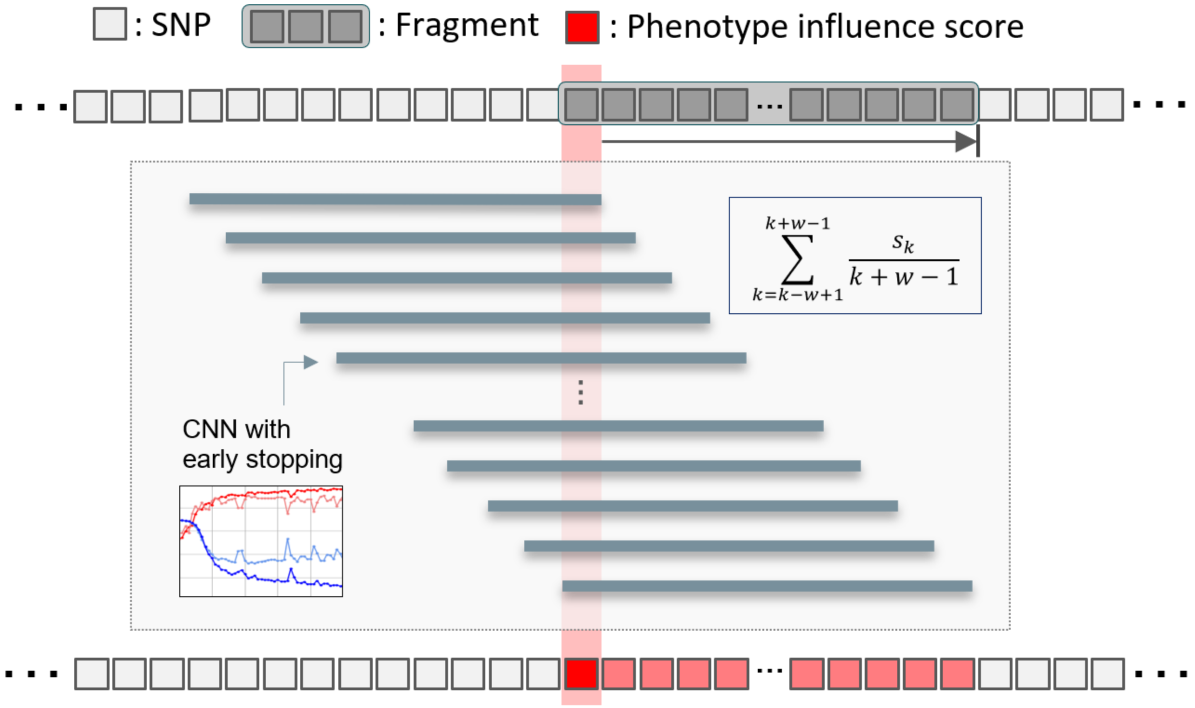
Prediction accuracy was calculated from deep learning applied to each fragment and converted to a z-score. The z-score follows a normal distribution with μ = 1 and σ = 0, under the hypothesis that there is no relationship between the variables in the population. Fragments with a z-score higher than the median were selected. An overlapping sliding window for the calculation of PIS is applied to these fragments (Figure 1). When the length of the fragment is w, the window is positioned w-1 from the first SNP of the fragment and moves by one SNP and stops at the last SNP of the fragment. Each region within the sliding window is divided into a train-test-validation set (60:20:20), and early stopping using a validation set is applied to prevent over-fitting. When the kth SNP is S_k, PIS is calculated as follows.

$$\sum_{k=k-w+1}^{k+w-1} \frac{s_k}{k+w-1}$$

This sliding window is applied to all selected fragments, resulting in a PIS score for all SNPs.

Phenotype classification using deep learning

We selected the top 100 to 10,000 SNPs based on the PIS. We used CNN, XG boost and Random Forest for the AD-CN classification with 10-fold cross validation. The CNN that we used consisted of convolution layer with a kernel size of 5, pooling lay with max-pool size of 2, fully connected layer of 64 nodes, and output layer with softmax activation function. XG Boost is a tree-based ensemble algorithm, one of popular implementations of gradient boosting. We trained XGboost using a “xgboost” package for python (https: //xgboost.readthedocs.io/). Random Forest is another ensemble learning method which uses many decision trees as its classifiers[56, 57]. We trained Random Forest using the scikit-learn package for python by setting the number of trees as 10 and the maximum depth of each tree as 3.


□ : SNP
: Fragment
🟥 : Phenotype influence score
$\sum_{k=k-w+1}^{k+w-1} \frac{S_k}{k+w-1}$
CNN with
early stopping

[Figure1] Framework to calculate phenotype influence scores of SNPs. We divided the whole genome into 134,955 fragments, each with 40 SNPs. To calculate a phenotype influence score for each of the 40 SNPs included in one fragment, we used an overlapping window approach and CNN. w is the number of SNPs in the fragment and S_k is the k_{th} SNP in the fragment.

Results

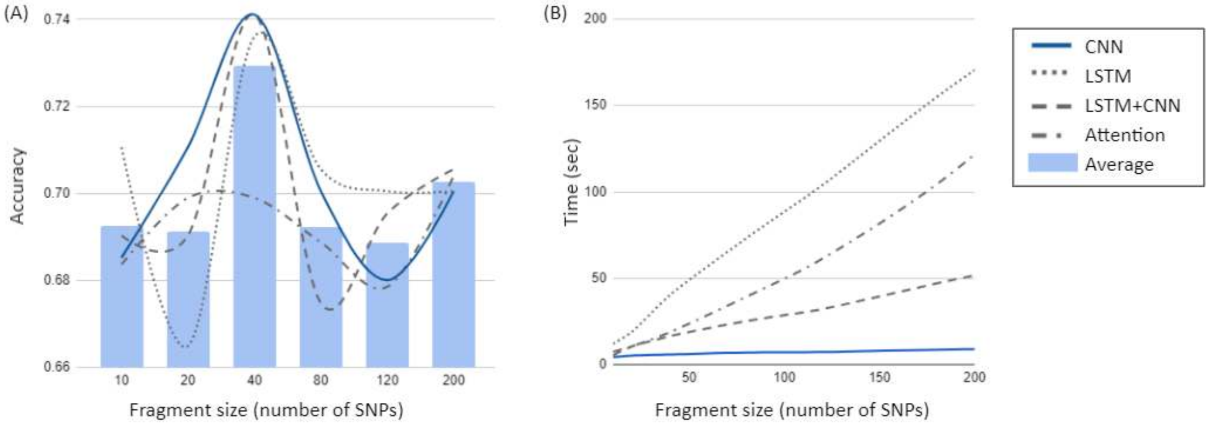
Our deep learning-based approach consists of three steps to select informative SNPs and develop an accurate classification model. In the first step, we divided the whole genome into non-overlapping fragments of an optimal size. To choose an optimal fragment size and an optimal deep learning algorithm, we calculated the mean accuracy and computation time for classification of AD using various fragment sizes containing 10 to 200 SNPs and several deep learning algorithms (CNN, LSTM, LSTM-CNN, Attention). In this analysis, we used 10-200 SNPs located within a region surrounding the *APOE* gene, the strongest and most robust AD genetic risk locus. Figure 2 showed the mean accuracy and computation time for CNN, LSTM, LSTM-CNN, and Attention as a function of the fragment size. As shown in Fig. 2A, the analysis yielded the highest accuracy for classification of AD for a fragment size with 40 SNPs (Fig. 2A). For the fragment with 40 SNPs within a region surrounding the *APOE* gene, both CNN and LSTM-CNN models had the highest accuracy for classification of AD, followed by LSTM. However, the computation time of CNN and LSTM models were 5.9 seconds and 40.4 seconds, respectively. The computation time of LSTM, LSTM-CNN, and Attention models sharply increased compared to CNN as the fragment contains more SNPs (Fig 2B). Thus, we chose the fragment of 40 SNPs and CNN as an optimal fragment size and an optimal deep learning algorithm, respectively. The whole genome was divided into

134,955 fragments, each with 40 SNPs. We ran CNN on each fragment to calculate z-scores based on classification accuracy and selected phenotype-associated fragments. We selected 1,802 fragments with z-scores higher than a median z-score.

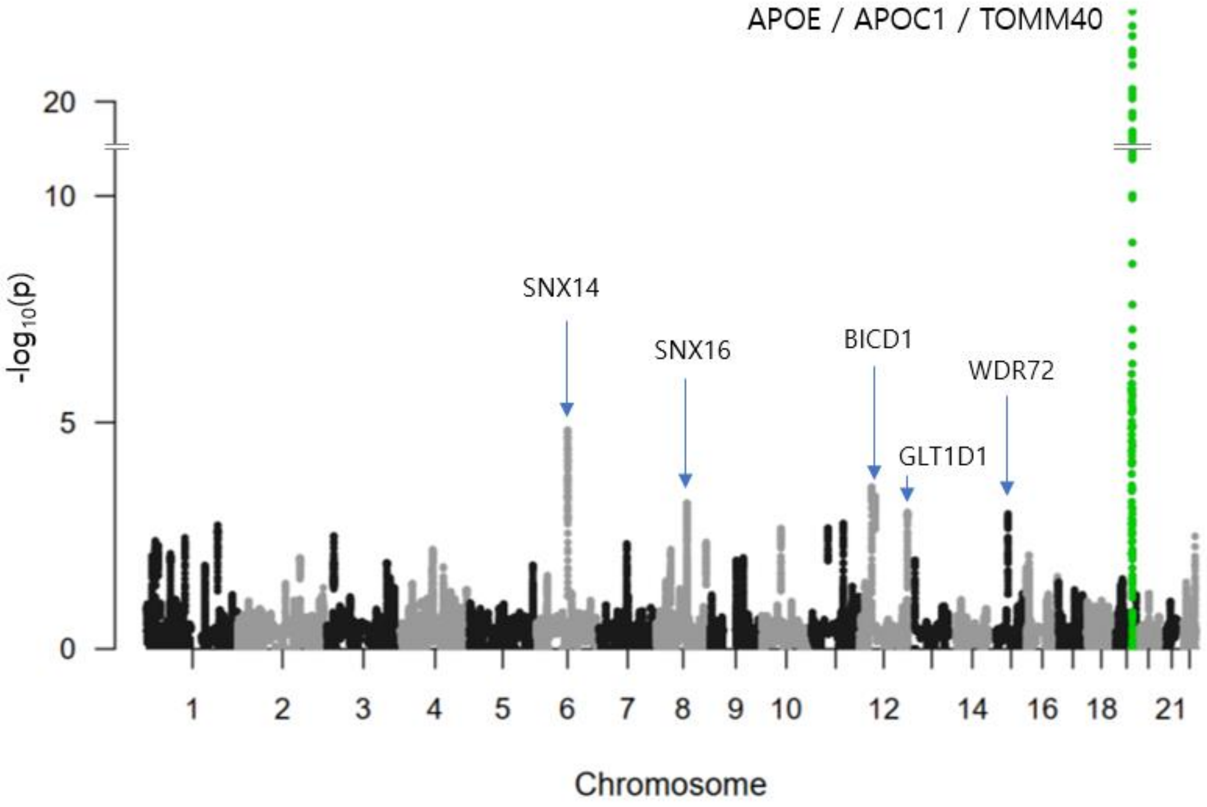
In the second step, using an overlapping window approach, we ran CNN on the selected fragments to calculate the PIS of each SNP in the selected fragments and identify phenotype-associated SNPs based on the PIS, as shown in Fig. 1. For each SNP, we calculated a mean accuracy of 40 windows, which is the PIS of the SNP. Using PIS values, we calculated the z-scores and one-tailed p-values. Figure 3 showed a Manhattan plot that plotted the -log10 p-values on the y-axis against the SNP position in the genome on the x-axis. The SNP with the smallest p-value was rs5117 in the *APOC1* gene (p-value=1.04E-22) and rs429358 in the *APOE* gene (p-value of 1.41E-16). The genetic region including *APOE*/*APOC1*/*TOMM40* genes is known as the strongest genetic risk locus for AD[58-61]. Next highest genetic loci were located at *SNX14*, *SNX16*, *BICD1*, *WDR72*, and *GLT1D1* genes.

In the third step, we ran CNN on the identified SNPs to develop an AD classification model. Table 1 shows the classification results of AD vs. CN using subsets containing the top 100 to 10,000 SNPs based on PIS. For comparison with traditional machine learning methods, we used two popular algorithms, XG Boost and Random Forest, as classifiers. The highest mean accuracy of 10-cross validation in classifying AD from CN by CNN was 75.02% (area under the curve (AUC) of 0.8157) for a subset containing 4,000 SNPs, which had 6.3% higher accuracy than Random Forest for a subset containing 2,000 SNPs and 1.94% higher accuracy than XG Boost for a subset containing 1,000 SNPs. When we calculated the classification accuracy of AD using only the number of *APOE* ε4 alleles, the classification accuracy was 66.7%, which was 8.3% lower than our method. Our CNN models outperformed two traditional machine learning models, Random Forest and XGBoost, in all cases as shown in Fig. 4.

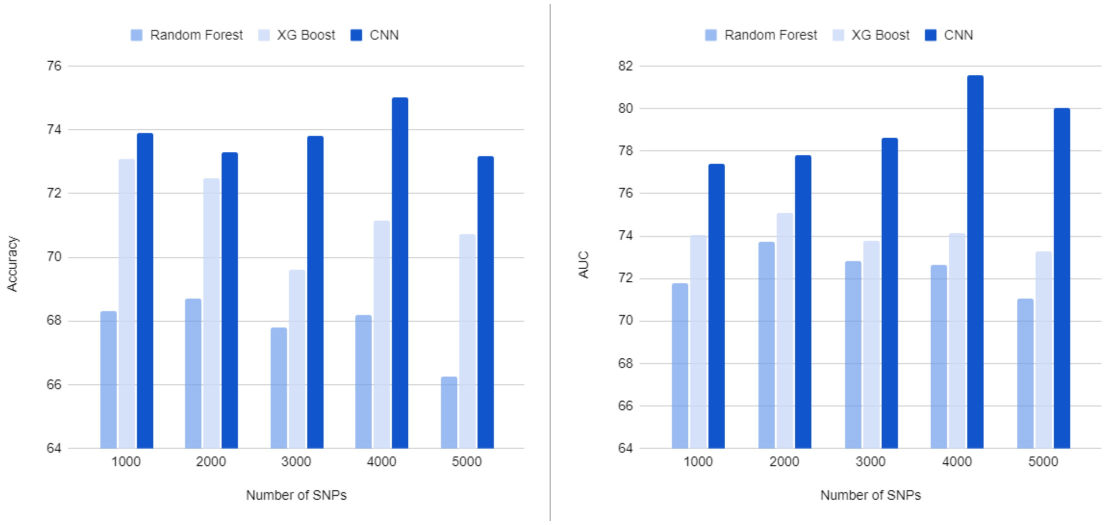
Figure 5 showed LocusZoom plots[62] for SNPs located at 300 kb upstream and downstream regions from the boundary of the *APOE* gene. The horizontal axis is the location of SNPs, and the vertical axis is -log10 of the p-values. Each dot represents a SNP and the color represents the squared correlation coefficient (r^2) with the most significant SNP. Figure 5A shows p-values calculated using PLINK and the most significant SNP was rs429358 in *APOE*. Figure 5B showed p-values calculated using our deep learning approach, and the most significant SNP was rs5117 in *APOC1*. In Fig. 5B, we can see a linear increase on the left side of rs5117 and a linear decrease on the right side of rs5117, which was different from PLINK results (Fig. 5A), which has no linear patterns. In addition, in Fig. 5B, we can see three strongly correlated SNPs (r^2>0.8) with rs5117 on the left side of rs5117 but no SNPs on the right side of rs5117.



[Figure2] Selection of an optimal fragment size and an optimal deep learning algorithm. In order to choose an optimal fragment size and an optimal deep learning algorithm, we calculated the mean accuracy and computation time for classification of AD using various fragment sizes containing 10 to 200 SNPs in the *APOE* region and several deep learning algorithms (CNN, LSTM, LSTM-CNN, and Attention). (A) Mean accuracy as a function of the fragment size. The highest accuracy for classification of AD was obtained with a fragment having 40 SNPs in CNN, LSTM-CNN and LSTM models. (B) Computation time as a function of the fragment size. The computation time of CNN and LSTM models are 5.9 seconds and 40.4 seconds, respectively. Especially the computation time of LSTM, LSTM-CNN, and Attention models sharply increases compared to CNN as the fragment contains more SNPs.



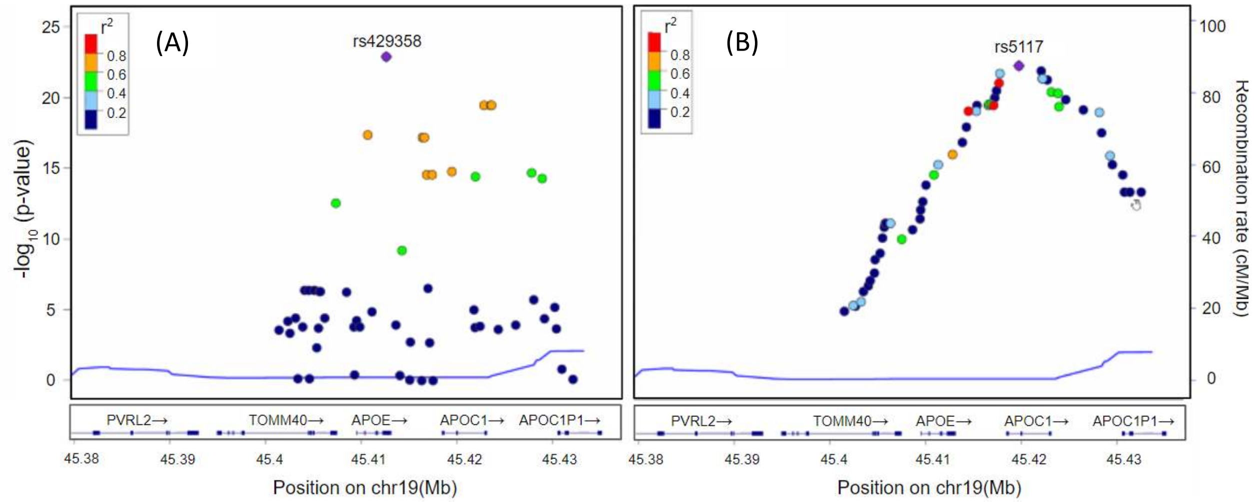
[Figure 3] Manhattan plot of p-values of SNPs by our deep learning based approach in AD. The X-axis shows SNP positions in the genome. The Y-axis shows -log10 of p-values. The genetic region including *APOE*, *APOC1*, and *TOMM40* genes is known as the strongest genetic risk locus for Alzheimer's disease. The SNP with the smallest p-value was rs5117 in *APOC1* gene (P=1.04E-22). rs429358 in *APOE* has a p-value of 1.41E-16. Next identified genetic loci were located at *SNX14*, *SNX16*, *BICD1*, *WDR72*, and *GLT1D1* genes.



[Figure 4] Results of classification of AD from CN. The X-axis shows the number of top SNPs selected based on phenotype influence score for AD classification. The Y-axis shows the accuracy (A) and AUC (B) of 10-fold cross-validation. Our CNN-based approach yielded the highest accuracy and AUC of 75.02% and 0.8157, respectively, for 4,000 SNPs. In all cases, our CNN models outperformed two traditional machine learning models, Random Forest and XG Boost.

Top	Random Forest				XG Boost				CNN			
	Accuracy	STD(±)	AUC	STD(±)	Accuracy	STD(±)	AUC	STD(±)	Accuracy	STD(±)	AUC	STD(±)
100	66.46	7.79	71.37	5.76	70.24	2.80	72.66	2.81	68.29	2.87	72.16	6.03
200	67.18	3.88	71.75	3.86	67.99	1.52	71.66	2.45	69.52	5.08	71.82	4.94
300	66.26	3.80	70.98	3.77	68.20	3.32	70.29	2.72	70.64	2.20	72.50	5.85
400	67.58	4.67	70.74	4.28	69.42	3.43	71.77	2.34	67.99	4.65	71.67	4.12
500	67.59	7.79	71.11	4.57	71.05	2.56	73.81	3.25	71.56	6.58	74.11	6.14
1000	68.31	5.22	71.78	4.45	**73.08**	2.89	74.07	3.72	73.91	3.87	77.41	4.44
2000	**68.70**	3.13	**73.72**	4.24	72.48	2.61	**75.09**	3.65	73.29	2.77	77.82	4.09
3000	67.78	3.59	72.82	3.51	69.62	4.27	73.76	3.28	73.80	2.40	78.62	2.82
4000	68.19	4.69	72.63	4.69	71.15	4.07	74.12	3.68	**75.02**	3.17	**81.57**	2.61
5000	66.25	5.41	71.05	3.99	70.74	3.14	73.30	3.05	73.19	4.72	80.03	5.06
10000	66.26	5.59	69.19	5.28	69.63	3.27	72.48	2.11	71.05	6.57	70.83	14.24

[Table 1] Results of classification of AD from CN. The table shows the number of top SNPs selected based on phenotype influence score for AD classification and the accuracy and AUC of 10-fold cross-validation. Our CNN-based approach yielded the highest accuracy and AUC of 75.02% and 0.8157, respectively, for 4,000 SNPs. In all cases, our CNN models outperformed two traditional machine learning models, Random Forest and XG Boost.



[Figure5] LocusZoom plots for SNPs located at the 300 kb upstream and downstream region from the boundary *APOE* gene. The horizontal axis is the location of SNPs and the vertical axis is -log10 of p-values. Each dot represents a SNP and the color represents the squared correlation coefficient (r^2) with the most significant SNP. (A) shows p-values calculated using PLINK and the most significant SNP was rs429358 in *APOE*. (B) shows p-values calculated using our deep learning approach and the most significant SNP was rs5117 in *APOC1*. In (B), we can see linear increase on the left side of rs5117 and linear decrease on the right side of rs5117, which was different from PLINK results (A), which has no linear patterns. In addition, in (B), we can see three strongly correlated SNPs (r^2>0.8) with rs5117 on the left side of rs5117 but no SNPs on the right side of rs5117.

Discussion

In this study, we propose a novel deep learning-based approach to select disease-associated SNPs and develop an accurate classification model using high dimensional genome data. We tested our approach using genome-wide genotyping data for Alzheimer's disease (AD) (N=981). The proposed method successfully identified significant genetic loci for AD that included the well-known AD genetic risk loci. The deep learning based approach outperformed traditional machine learning methods for classification of AD.

The deep learning-based approach for identification of genetic variants consists of three steps. In the first step, we divided the whole genome into non-overlapping fragments with an optimal size. Although deep learning has solved many real-world problems, due to the high dimensionality of the genomic data, few deep learning approaches have been used in GWAS to identify genetic variants and disease classification [22]. To our knowledge, this study is the first to develop a deep learning-based method for identifying genetic variants by using a fragmentation and windowing approach.

In the second step, we calculated a PIS of each SNP within the selected fragments by using an overlapping window and CNN algorithm. Our method calculates PIS, a novel index which is used to find disease-related variants and predict disease. Furthermore, we calculated the z-scores and one-tailed p-values using PIS, which yielded a Manhattan plot showing the most significant genetic loci in *APOE*/*APOC1*/*TOMM40* genes that are known as the strongest genetic risk factors for AD. Our method also identified several novel candidate genetic loci. Sorting nexin (SNX) *14* and *SNX16* on chromosomes 6 and 8, respectively, have not been previously identified to be associated with AD though there may be special relevance for neurodegeneration as *SNX12*[63], *SNX17*[64], *SNX27*[65], and *SNX33*[66] are involved in neuronal survival. Bicaudal D1 (*BICD1*) on chromosome 12 is a susceptibility gene in chronic obstructive pulmonary disease[67] and lissencephaly[68], but there are no reports of it being associated with AD.

In the third step, we selected top SNPs based on PIS to develop classification models for AD. We selected sets of highly AD-related SNPs, and classified AD from CN using CNN and two popular traditional machine learning algorithms, XGBoost and Random Forest. We found the accuracy of classification was changed with the number of the selected SNPs and the classification algorithms. The highest mean accuracy of the classification was 75.0% when CNN was used on the top 4,000 SNPs, which outperformed two traditional machine learning algorithms. It was also 8.3% higher than the accuracy of the classification using only the number of *APOE* ε4 alleles. Classification is the first step toward achieving a better understanding of the genetic architecture of AD. The proposed method will benefit from future studies that use deep learning with quantitative phenotypes and baseline values to predict future disease trajectories.

We plotted the SNPs selected by PIS and PLINK for comparison using LocusZoom. We found that there were no SNPs with r^2 greater than 0.8 in the PLINK results, but three strongly associated SNPs were identified using our method. This is because the PLINK method finds statistical significance SNP by SNP, while the approach of deep learning uses multiple inputs to adjust weights through the training process. Deep learning uses adjacent SNPs to compute gradients at every epoch and uses a loss function to adjust the weights in the backpropagation process. Unlike PLINK, our method shows that SNPS related to phenotype can be extracted by considering surrounding SNPs, which means that both methods might be complementary because they identify different variants though notably in the same region around APOE.

In summary, our novel deep learning-based approach can identify AD-related SNPs by using genome-wide data and develop a classification model for AD. The heritability of AD is estimated to be up to 80%. Accordingly, it is important to identify novel genetic loci related to the disease. Using a modest sample size, we found a significant genetic locus and a classification accuracy of 75%. In future work, we plan to apply our method to large-scale whole genome sequencing data sets that are expected to become available soon to identify novel AD-related SNPs and develop more accurate classification models. We also plan to study early stages of disease including mild cognitive impairment and relate variation to quantitative endophenotypes that may be more informative than binary classification.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01AG024904) and DODADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Author Contributions

TJ, KN, and AS: Conceptualization. AS: Acquisition of Data and Interpretation of Results. TJ, KN: Data Curation. TJ: Formal Analysis, Investigation, Methodology, Validation, Visualization. TJ, KN, and AS: Writing - Original Draft Preparation. TJ, KN, PB and AS: Review & Editing.

Funding

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). In addition, this work was supported, in part, by grants from the National Institutes of Health (NIH) and includes the following sources: P30 AG010133, P30 AG072976, R01 AG019771, R01 AG057739, U01 AG024904, R01 LM013463, R01 AG068193, T32 AG071444, U01 AG068057, R01 LM012535 and R03 AG063250. The funders played no role in the design of the study, analysis, and interpretation of the data or in writing the manuscript.

Reference

1. Wainberg M, Merico D, Delong A, Frey BJ. Deep learning in biomedicine. Nature Biotechnology 2018;36:829-838.
2. Jo T, Nho K, Saykin AJ. Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. Frontiers in aging neuroscience 2019;11:220.
3. Jo T, Nho K, Risacher SL, Saykin AJ, for the Alzheimer's Neuroimaging I. Deep learning detection of informative features in tau PET for Alzheimer's disease classification. BMC Bioinformatics 2020;21:496.
4. Zhang Z, Park CY, Theesfeld CL, Troyanskaya OG. An automated framework for efficiently designing deep convolutional neural networks in genomics. Nature Machine Intelligence 2021.
5. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods 2015;12:931-934.
6. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology 2015;33:831-838.
7. Xiong HY, Alipanahi B, Lee LJ, et al. The human splicing code reveals new insights into the genetic determinants of disease. Science 2015;347:1254806.
8. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Research 2016;44:e107-e107.
9. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biology 2017;18:67.
10. Zhang S, Hu H, Jiang T, Zhang L, Zeng J. TITER: predicting translation initiation sites by deep learning. Bioinformatics 2017;33:i234-i242.
11. Tasaki S, Gaiteri C, Mostafavi S, Wang Y. Deep learning decodes the principles of differential gene expression. Nature Machine Intelligence 2020;2:376-386.
12. Zheng A, Lamkin M, Zhao H, Wu C, Su H, Gymrek M. Deep neural networks identify sequence context features predictive of transcription factor binding. Nature Machine Intelligence 2021;3:172-180.
13. Scherer M, Schmidt F, Lazareva O, et al. Machine learning for deciphering cell heterogeneity and gene regulation. Nature Computational Science 2021;1:183-191.
14. Listgarten J, Weinstein M, Kleinstiver BP, et al. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. Nature Biomedical Engineering 2018;2:38-47.
15. Shen MW, Arbab M, Hsu JY, et al. Predictable and precise template-free CRISPR editing of pathogenic variants. Nature 2018;563:646-651.
16. Leenay RT, Aghazadeh A, Hiatt J, et al. Large dataset enables prediction of repair after CRISPR–Cas9 editing in primary T cells. Nature Biotechnology 2019;37:1034-1037.
17. Liu Q, He D, Xie L. Prediction of off-target specificity and cell-specific fitness of CRISPR-Cas System using attention boosted deep learning and network-based gene feature. PLOS Computational Biology 2019;15:e1007480.
18. Kim HK, Min S, Song M, et al. Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity. Nature Biotechnology 2018;36:239-241.
19. Ogden PJ, Kelsic ED, Sinai S, Church GM. Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design. Science 2019;366:1139.
20. Yan J, Qiu Y, Ribeiro dos Santos AM, et al. Systematic analysis of binding of transcription factors to noncoding variants. Nature 2021;591:147-151.
21. Buniello A, MacArthur JA L, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research 2019;47:D1005-D1012.

22. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. Nature 2015;526:68-74.
23. Li F, Yang Y, Xing EP. From Lasso regression to feature vector machine. Proceedings of the 18th International Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada: MIT Press, 2005: 779–786.
24. Yamada M, Jitkrittum W, Sigal L, Xing EP, Sugiyama M. High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso. Neural Computation 2014;26:185-207.
25. Xu Z, Huang G, Weinberger KQ, Zheng AX. Gradient boosted feature selection. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, New York, USA: Association for Computing Machinery, 2014: 522–531.
26. Amaldi E, Kann V. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. Theoretical Computer Science 1998;209:237-260.
27. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–1182.
28. Veitch DP, Weiner MW, Aisen PS, et al. Understanding disease progression and improving Alzheimer's disease clinical trials: Recent highlights from the Alzheimer's Disease Neuroimaging Initiative. Alzheimers Dement 2019;15:106-152.
29. Saykin AJ, Shen L, Yao X, et al. Genetic studies of quantitative MCI and AD phenotypes in ADNI: Progress, opportunities, and plans. Alzheimers Dement 2015;11:792-814.
30. Saykin AJ, Shen L, Yao X, et al. Genetic studies of quantitative MCI and AD phenotypes in ADNI: Progress, opportunities, and plans. Alzheimers Dement 2015;11:792-814.
31. Park YH, Hodges A, Risacher SL, et al. Dysregulated Fc gamma receptor-mediated phagocytosis pathway in Alzheimer's disease: network-based gene expression analysis. Neurobiol Aging 2020;88:24-32.
32. Horgusluoglu-Moloch E, Nho K, Risacher SL, et al. Targeted neurogenesis pathway-based gene analysis identifies ADORA2A associated with hippocampal volume in mild cognitive impairment and Alzheimer's disease. Neurobiol Aging 2017;60:92-103.
33. Freedman ML, Reich D, Penney KL, et al. Assessing the impact of population stratification on genetic association studies. Nat Genet 2004;36:388-393.
34. Park YH, Hodges A, Simmons A, et al. Association of blood-based transcriptional risk scores with biomarkers for Alzheimer disease. Neurol Genet 2020;6:e517.
35. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-575.
36. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems; 2012: 1097-1105.
37. Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation 1997;9:1735-1780.
38. Zhang J, Li Y, Tian J, Li T. LSTM-CNN Hybrid Model for Text Classification. 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC); 2018 12-14 Oct. 2018: 1675-1680.
39. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv preprint arXiv:170603762 2017.
40. Rosenblatt F. The perceptron, a perceiving and recognizing automaton Project Para: Cornell Aeronautical Laboratory, 1957.
41. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review 1958;65:386.
42. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics 1943;5:115-133.

43. Widrow B, Hoff ME. Adaptive switching circuits: Stanford Univ Ca Stanford Electronics Labs, 1960.

44. Minsky M, Papert SA. Perceptrons: An introduction to computational geometry: MIT press, 2017.

45. Werbos PJ. Applications of advances in nonlinear sensitivity analysis. System modeling and optimization: Springer, 1982: 762-770.

46. Werbos PJ. Backwards differentiation in AD and neural nets: Past links and new opportunities. Automatic differentiation: Applications, theory, and implementations 2006:15-34.

47. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. nature 1986;323:533-536.

48. LeCun Y, Touresky D, Hinton G, Sejnowski T. A theoretical framework for back-propagation. Proceedings of the 1988 connectionist models summer school; 1988: 21-28.

49. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning: MIT press Cambridge, 2016.

50. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. Icml; 2010.

51. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. Proceedings of the fourteenth international conference on artificial intelligence and statistics; 2011: JMLR Workshop and Conference Proceedings: 315-323.

52. Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. Journal of machine learning research 2011;12.

53. Hinton G, Srivastava N, Swersky K. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. Cited on 2012;14.

54. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980 2014.

55. Sutskever I, Martens J, Dahl G, Hinton G. On the importance of initialization and momentum in deep learning. International conference on machine learning; 2013: PMLR: 1139-1147.

56. Breiman L. Random Forests. Machine Learning 2001;45:5-32.

57. Jo T, Cheng J. Improving protein fold recognition by random forest. BMC Bioinformatics 2014;15:S14.

58. Saunders AM, Strittmatter WJ, Schmechel D, et al. Association of apolipoprotein E allele ϵ4 with late-onset familial and sporadic Alzheimer's disease. Neurology 1993;43:1467-1467.

59. Roses AD, Lutz MW, Amrine-Madsen H, et al. A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease. The Pharmacogenomics Journal 2010;10:375-384.

60. Corder E, Saunders A, Strittmatter W, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 1993;261:921-923.

61. Cervantes S, Samaranch L, Vidal-Taboada JM, et al. Genetic variation in APOE cluster region and Alzheimer's disease risk. Neurobiology of Aging 2011;32:2107.e2107-2107.e2117.

62. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 2010;26:2336-2337.

63. Zhao Y, Wang Y, Yang J, et al. Sorting nexin 12 interacts with BACE1 and regulates BACE1-mediated APP processing. Molecular neurodegeneration 2012;7:1-10.

64. Lee J, Retamal C, Cuitiño L, et al. Adaptor protein sorting nexin 17 regulates amyloid precursor protein trafficking and processing in the early endosomes. Journal of Biological Chemistry 2008;283:11501-11508.

65. Gallon M, Clairfeuille T, Steinberg F, et al. A unique PDZ domain and arrestin-like fold interaction reveals mechanistic details of endocytic recycling by SNX27-retromer. Proceedings of the National Academy of Sciences 2014;111:E3604-E3613.

66. Heiseke A, Schöbel S, Lichtenthaler SF, et al. The Novel Sorting Nexin SNX33 Interferes with Cellular PrPSc Formation by Modulation of PrPc Shedding. Traffic 2008;9:1116-1129.

67. Mercado N, Colley T, Baker JR, et al. Bicaudal D1 impairs autophagosome maturation in chronic obstructive pulmonary disease. FASEB BioAdvances 2019;1:688-705.

68. Swan A, Nguyen T, Suter B. Drosophila Lissencephaly-1 functions with Bic-D and dynein in oocyte determination and nuclear positioning. Nature Cell Biology 1999;1:444-449.