African Journal of Mathematics and Statistics Studies

ISSN: 2689-5323

Volume 7, Issue 2, 2024 (pp. 112-130)

www.abjournals.org

APPLICATION OF DEEP LEARNING FOR THE DETECTION OF GENETIC VARIATIONS: ITS IMPLEMENTATION IN CLASSIFYING ALZHEIMER'S DISEASE

Ugwuanyi Ifesinachi[1], Oladoyin Idris Atolagbe[2], Anazor Chinenye[3],

Dike Ikechukwu[4], Ezulu Priscilla Chinwendu[5], Nwagbata Amarachukwu[6]

[1-6]Department of Statistics, University of Nigeria Nsukka.

Cite this article:

Ugwuanyi I., Oladoyin I. A., Anazor C., Dike I., Ezulu P. C., Nwagbata A. (2024), Application of Deep Learning for the Detection of Genetic Variations: Its Implementation in Classifying Alzheimer's Disease. African Journal of Mathematics and Statistics Studies 7(2), 112-130. DOI: 10.52589/AJMSS-4WNIT6F9

Manuscript History

Received: 13 Jan 2024

Accepted: 25 Mar 2024

Published: 26 Apr 2024

ABSTRACT: *Deep learning emerges as a promising technique, utilizing nonlinear transformations for feature extraction from high-dimensional datasets. However, its application encounters challenges in genome-wide association studies (GWAS) dealing with high-dimensional genomic data. This study introduces an innovative three-step method termed SWAT-CNN for the identification of genetic variants. This approach employs deep learning to pinpoint phenotype-related single nucleotide polymorphisms (SNPs), facilitating the development of precise disease classification models. In the first step, the entire genome undergoes division into non overlapping fragments of an optimal size. Subsequently, convolutional neural network (CNN) analysis is conducted on each fragment to identify phenotype-associated segments. The second step, employs a Sliding Window Association Test (SWAT), where CNN is utilized on the selected fragments to compute phenotype influence scores (PIS) and detect phenotype-associated SNPs based on these scores. The third step involves running CNN on all identified SNPs to construct a comprehensive classification model. Validation of the proposed approach utilized GWAS data from the Alzheimer's disease Neuroimaging Initiative (ADNI), encompassing 981 subjects, including cognitively normal older adults (CN) and individuals with Alzheimer's disease (AD). Notably, the method successfully identified the widely recognized APOE region as the most significant genetic locus for AD. The resulting classification model exhibited an area under the curve (AUC) of 0.82, demonstrating compatibility with traditional machine learning approaches such as random forest and XGBoost. SWAT-CNN, as a groundbreaking deep learning-based genome-wide methodology, not only identified AD-associated SNPs but also presented a robust classification model for Alzheimer's disease, suggesting potential applications across diverse biomedical domains.*

KEYWORDS: Deep neural networks, genetic variations, Alzheimer's disease, GWAS, phenotype impact scores.

Article DOI: 10.52589/AJMSS-4WNIT6F9
DOI URL: https://doi.org/10.52589/AJMSS-4WNIT6F9

INTRODUCTION

Deep learning, a prominent machine learning algorithm, stands out for its capability to perform nonlinear transformations for extracting features from high-dimensional data [1]. This is in contrast to traditional machine learning models, which predict a linear combination of weights by assuming a linear relationship between input features and the phenotype of interest. In the medical field, deep learning has proven to be effective in predicting disease outcomes by directly handling original high-dimensional medical imaging data without the need for feature selection procedures [2, 3]. In genetic research, deep learning frameworks have been applied to explore molecular phenotypes predicting the effects of noncoding variants [4–10], differential gene expression [11], and potential transcription factor binding sites [12]. These frameworks utilize CHIP-Seq or DNase-Seq data as training data to predict chromatin features from DNA sequences.

Recent applications of deep learning extend to capturing mutations and analyzing gene regulations, showcasing its potential in advancing our understanding of epigenetic regulation [13]. Moreover, in the field of gene therapy, deep learning is actively employed to design CRISPR guide RNAs using gene features derived from deep learning models [14–19].

Genome-wide association studies (GWAS) conventionally use a statistical approach, considering one single nucleotide polymorphism (SNP) at a time across the entire genome to identify population-based genetic risk variations for human diseases and traits [20, 21]. However, deep learning has yet to be extensively employed in GWAS due to the inherent challenges posed by the high-dimension low-sample-size (HDLSS) problem [22], which significantly impacts phenotype prediction using genetic variation. Although feature reduction approaches are commonly applied [23–25], resolving this problem remains challenging, particularly when dealing with high-dimensional genomic data. Therefore, there is a need to develop a deep learning framework specifically tailored for identifying genetic variants using whole-genome data.

This study introduces a novel three-step deep learning-based approach to select informative SNPs and develop classification models for a target phenotype. In the initial step, the whole genome is divided into non overlapping fragments of an optimal size, and deep learning algorithms are utilized to select phenotype-associated fragments containing relevant SNPs. various fragment sizes and deep learning algorithms are tested to determine the optimal combination. The second step involves running the optimal deep learning algorithm using an overlapping Sliding Window Association Test (SWAT) within selected fragments to calculate phenotype influence scores (PIS) using SNPs and the target phenotype, thus identifying informative SNPs. The final step consists of running the optimal algorithm on all identified informative SNPs to construct a comprehensive classification model. Focusing on Alzheimer's disease (AD), the most prevalent form of dementia, characterized by progressive memory and cognitive function deterioration, this study tested the proposed approach using only whole-genome data for AD (N = 981; cognitively normal older adults (CN) = 650 and AD = 331). The approach successfully identified the well-known APOE region as the most significant genetic locus for AD. Utilizing this identified region, a classification model was developed using Convolutional Neural Network (CNN). To assess the algorithm's performance relative to traditional machine learning algorithms, XGBoost and random forest were also applied. The classification model achieved 75.2% accuracy, showing compatibility with traditional machine learning methods and outperforming XGBoost and random forest by 3.8% and 9.6%, respectively. This novel deep learning-based approach demonstrates the potential to identify informative SNPs and develop an accurate classification model for AD by aggregating nearby SNPs.

Materials and Method

Sourcing the Datasets

The individuals analyzed in this study were sourced from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort [53, 54]. The ADNI was initiated with its first phase (ADNI-1) in 2003, with the objective of evaluating the viability of integrating serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, as well as clinical and neuropsychological assessments to assess the progression of mild cognitive impairment (MCI) and early-stage Alzheimer's disease (AD). Subsequent phases (ADNI-GO, ADNI-2, and ADNI-3) extended ADNI-1 for continuous follow-up of existing participants and the inclusion of new enrollees. Comprehensive demographic information, APOE and whole genome genotyping data, and clinical details are openly accessible via the ADNI data repository (www.loni.usc.edu/ADNI/). Informed consent was obtained from all subjects.

Top of Form

Genotype and Imputation

ADNI participants underwent genotyping utilizing various Illumina platforms, including Illumina Human610-Quad BeadChip, Illumina HumanOmniExpress BeadChip, and Illumina HumanOmni 2.5M BeadChip [54]. Due to the diverse genotyping platforms employed by ADNI, distinct quality control procedures were independently applied to each genotyping platform's data. Imputation of un-genotyped single nucleotide polymorphisms (SNPs) was carried out separately using MACH and the Haplotype Reference Consortium (HRC) data as a reference panel [55].

Before the imputation process, rigorous quality control measures were implemented for both samples and SNPs, consistent with established criteria: (1) SNP criteria included SNP call rate < 95%, Hardy–Weinberg P value < 1 × 10−6, and minor allele frequency (MAF) < 1%, and (2) sample criteria comprised sex inconsistencies and sample call rate < 95% [56]. Additionally, to mitigate spurious associations arising from population stratification, only non-Hispanic participants of European ancestry, aligning with HapMap CEU (Utah residents with Northern and Western European ancestry from the CEPH collection) or TSI (Toscani in Italia) populations based on multidimensional scaling (MDS) analysis and HapMap genotype data, were selected [56, 57].

Post-imputation, standard quality control procedures were applied to the imputed genotype data, consistent with previously established protocols [58]. Notably, an r2 value of 0.30 served as the threshold for accepting imputed genotypes. In this study, imputed genome-wide genotyping data from 981 ADNI non-Hispanic participants (650 cognitively normal older adults (CN) and 331 AD patients) were utilized, encompassing a total of 5,398,183 SNPs (minor allele frequency (MAF) > 5%).

Article DOI: 10.52589/AJMSS-4WNIT6F9
DOI URL: https://doi.org/10.52589/AJMSS-4WNIT6F9

Genome-wide Association Study (GWAS)

Utilizing imputed genotypes, a genome-wide association study (GWAS) for Alzheimer's disease (AD) was undertaken. Logistic regression, incorporating age and sex as covariates, was executed using PLINK [59] to assess the association of each single nucleotide polymorphism (SNP) with AD. To account for the impact of multiple testing, a stringent threshold for genome-wide significant association ($P < 5 \times 10-8$) was applied, utilizing a Bonferroni correction.

Fragmentation of Whole Genome Data

Genomic information for 981 participants was partitioned into non-overlapping fragments, ranging in size from 10 SNPs to 200 SNPs, with the aim of identifying the optimal fragment size. These subsets, each comprising fragments of the same size, were further segmented into train–test–validation sets in a ratio of 60:20:20. Convolutional neural network (CNN) [60], long short-term memory (LSTM) [61], LSTM-CNN [62], and attention [63] algorithms were individually applied to each subset. To prevent overfitting, early stopping mechanisms were implemented based on a validation set, followed by the assessment of training time and accuracy (ACC).

Deep Learning on Fragments

The evolution of deep learning can be traced back to seminal developments such as the perceptron [64, 65], which introduced weight adjustment to mimic human brain behavior through interconnected neurons with on–off functions in a network structure [66], and Adaline [67], which utilized gradient descent for weight updates. These early neural networks progressed into the multilayer perceptron, incorporating hidden layers to address the XOR problem [68]. A pivotal theoretical advancement occurred with the introduction of backpropagation to update hidden layer weights [69–72]. Overcoming the inherent challenge of vanishing gradients in backpropagation, particularly in deep networks [73], was addressed through the incorporation of activation functions like the sigmoid and ReLU [74, 75], and the development of optimization methods enhancing gradient descent, such as Ada-Grad [76], RMSprop [77], and Adam [78]. These advancements, coupled with GPU hardware improvements, ushered in the era of contemporary deep learning.

Deep learning has established the theoretical underpinnings of backpropagation, activation functions, and optimization methods for enhanced gradient descent. Widely used deep learning algorithms, including CNN, LSTM, and attention, embody a hierarchical structure that refines and extends the foundational principles of deep learning. The intricate technical details of each algorithm are extensively expounded in relevant literature; hence, our focus here centers on the core deep learning technology applied in our experiment.

In our experiments, we employed ReLU as the activation function underlying the deep learning algorithms. ReLU, a prevalent choice in the deep learning community, replaces values less than zero with zero and retains values greater than zero. This characteristic ensures a derivative of one when the value is positive, facilitating gradient adjustments through the hidden layer without vanishing. The optimization method chosen was Adam, currently the most favored for deep learning. Adam leverages momentum SGD [79] and RMSprop, expressed as Gt being the sum of the squared modified gradient, and ε is a constant preventing division by zero in the equation.

Article DOI: 10.52589/AJMSS-4WNIT6F9
DOI URL: https://doi.org/10.52589/AJMSS-4WNIT6F9

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

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

$$\hat{V}_t = \frac{V_t}{1 - \gamma_1^t}$$

$$\hat{G}_t = \frac{G_t}{1 - \gamma_2^t}$$

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

Backpropagation is employed to compute the initial error value based on a randomly assigned weight, utilizing the least squares method. Subsequently, it iteratively updates the weight through the chain rule until the derivative becomes zero. In this context, a derivative value of zero signifies that the weight remains unchanged when the gradient is subtracted from the preceding weight.

$$W_o(t+1) = W_o t - \frac{\partial \text{Error} Y_o}{\partial W_o}$$

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

When yo1 and yo2 represent the output values from the output layer passing through the hidden layer, and the actual values of the given data are denoted as yt1 and yt2, the partial derivative of the error (ErrorYo) with respect to the weight of the output layer can be computed using the chain rule as follows:

Article DOI: 10.52589/AJMSS-4WNIT6F9
DOI URL: https://doi.org/10.52589/AJMSS-4WNIT6F9

$$\frac{\partial \mathrm{Error} Y_{\mathrm{o}}}{\partial w_{\mathrm{o}}}=\frac{\partial \mathrm{Error} Y_{\mathrm{o}}}{\partial y_{\mathrm{o}1}} \cdot \frac{\partial y_{\mathrm{o}1}}{\partial \mathrm{net3}} \cdot \frac{\partial \mathrm{net3}}{\partial w_{\mathrm{o}}}$$

The calculation for the partial derivative of the error (ErrorYo) with respect to the weight of the hidden layer is as follows:

$$\frac{\partial \mathrm{Error} Y_{\mathrm{o}}}{\partial \mathrm{h}_1}=\frac{\partial\left(\mathrm{Error} y_{\mathrm{o}1}+\mathrm{Error} y_{\mathrm{o}2}\right)}{\partial y_{\mathrm{h}1}}=\frac{\frac{\partial \mathrm{Error} y_{\mathrm{o}1}}{\partial y_{\mathrm{h}1}}}{(\mathrm{a})}+\frac{\frac{\partial \mathrm{Error} y_{\mathrm{o}2}}{\partial y_{\mathrm{h}1}}}{(\mathrm{b})}$$

(a)

$$\begin{aligned}\frac{\partial \mathrm{Error} y_{\mathrm{o}1}}{\partial y_{\mathrm{h}1}} &=\frac{\partial \mathrm{Error} y_{\mathrm{o}1}}{\partial \mathrm{net}_3} \cdot \frac{\partial \mathrm{net}_3}{\partial y_{\mathrm{h}1}} \\ &=\left(y_{\mathrm{o}1}-y_{t1}\right) y_{\mathrm{o}1}\left(1-y_{\mathrm{o}1}\right) y_{\mathrm{o}1}\end{aligned}$$

(b)

$$\begin{aligned}\frac{\partial \mathrm{Error} y_{\mathrm{o}2}}{\partial y_{\mathrm{h}1}} &=\frac{\partial \mathrm{Errory}_{\mathrm{o}2}}{\partial \mathrm{net}_4} \cdot \frac{\partial \mathrm{net}_4}{\partial y_{\mathrm{h}1}} \\ &=\left(y_{\mathrm{o}2}-y_{t2}\right) y_{\mathrm{o}2}\left(1-y_{\mathrm{o}2}\right) y_{\mathrm{o}2}\end{aligned}$$

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

$$\frac{\partial \mathrm{Error} Y_{\mathrm{o}}}{\partial w_{\mathrm{h}}}=\frac{\partial \mathrm{Error} Y_{\mathrm{o}}}{\partial y_{\mathrm{h}1}} \cdot \frac{\partial y_{\mathrm{h}1}}{\partial \mathrm{net}_1 y} \cdot \frac{\partial \mathrm{net}_1}{\partial w_{\mathrm{h}}}$$

$$=\left(\delta y_{\mathrm{o}1} y_{\mathrm{o}1}-\delta y_{\mathrm{o}2} y_{\mathrm{o}2}\right) y_{\mathrm{h}1}\left(1-y_{\mathrm{h}1}\right) x_1$$

Calculation of Phenotype Influence Score Using Deep Learning

Predictive accuracy was determined through the application of deep learning to each fragment and then transformed into a z-score. The z-score conforms to a normal distribution with parameters μ = 1 and σ = 0, assuming no intrinsic relationship between the variables in the population. Fragments with z-scores exceeding the median were chosen for further analysis.

An overlapping Sliding Window Association Test (SWAT) was subsequently employed for calculating Phenotype Influence Scores (PIS) within these selected fragments. In this process, with the fragment length denoted as 'w,' the window is positioned at 'w-1' from the initial SNP of the fragment, progressing by one SNP until reaching the final SNP of the fragment. Each region within the SWAT undergoes subdivision into a train–test–validation set (60:20:20), with the inclusion of early stopping using a validation set to prevent overfitting. When the kth SNP is represented as Sk, the calculation of PIS is carried out as follows:

Top of Form

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

The Sliding Window Association Test (SWAT) is systematically applied to all chosen fragments, generating Phenotype Influence Scores (PIS) for all Single Nucleotide Polymorphisms (SNPs).

Phenotype Classification Using Deep Learning

We opted for the top 100 to 10,000 SNPs based on Phenotype Influence Scores (PIS). For the classification of AD-CN, a Convolutional Neural Network (CNN) was utilized, incorporating convolution layers with a kernel size of 5, pooling the layer with a max-pool size of 2, a fully connected layer comprising 64 nodes, and an output layer with a softmax activation function. The challenges posed by gradient vanishing and explosion, resulting from the recurrent weight matrix's repeated multiplication, hindered the effective training of Recurrent Neural Network (RNN) or its variants. To facilitate performance comparison, we also applied traditional tabular data classification algorithms such as Random Forest and XGBoost. XGBoost, a popular gradient boosting implementation, was trained using the 'xgboost' package for Python (https://xgboost.readthedocs.io/). Random Forest, an ensemble learning method utilizing multiple decision trees as classifiers [80, 81], was trained using the scikit-learn package for Python, with the number of trees set to 10 and the maximum depth of each tree limited to 3.

RESULTS

Our deep learning approach comprises three sequential steps aimed at the identification of informative SNPs and the construction of an accurate classification model. In the initial step, we partitioned the entire genome into nonoverlapping fragments of an optimized size. To determine the optimal fragment size and the most suitable deep learning algorithm, we assessed the mean accuracy and computation time for Alzheimer's disease (AD) classification across various fragment sizes containing 10 to 200 SNPs, employing several deep learning algorithms (CNN, LSTM, LSTM-CNN, Attention). The analysis specifically focused on 10–200 SNPs within a region surrounding the APOE gene, recognized as the most potent and resilient genetic risk locus for AD. Figure 1 illustrates the average accuracy and computation time for CNN, LSTM, LSTM-CNN, and attention as functions of the fragment size.

As depicted in Figure 1A, the analysis revealed the highest accuracy for AD classification with a fragment size of 40 SNPs. Figure 1B demonstrates the average accuracy and time concerning the deep learning algorithm, with a window size of 40 within the APOE gene region. CNN and LSTM-CNN models exhibited the highest accuracy for AD classification, followed by LSTM. However, the computation time for CNN and LSTM models was 5.9 and 40.4 seconds, respectively. The computation time of LSTM, LSTM-CNN, and attention models significantly increased compared to CNN due to the larger number of SNPs in the fragment. Consequently, we selected a fragment with 40 SNPs as the optimal fragment size for CNN and the optimal deep learning algorithm. The entire genome was subsequently divided into 134,955 fragments, each containing 40 SNPs. We applied CNN to each fragment, calculating z-scores based on classification accuracy, and identified phenotype-associated fragments. A total of 1,802 fragments with z-scores surpassing the median were chosen for further analysis.

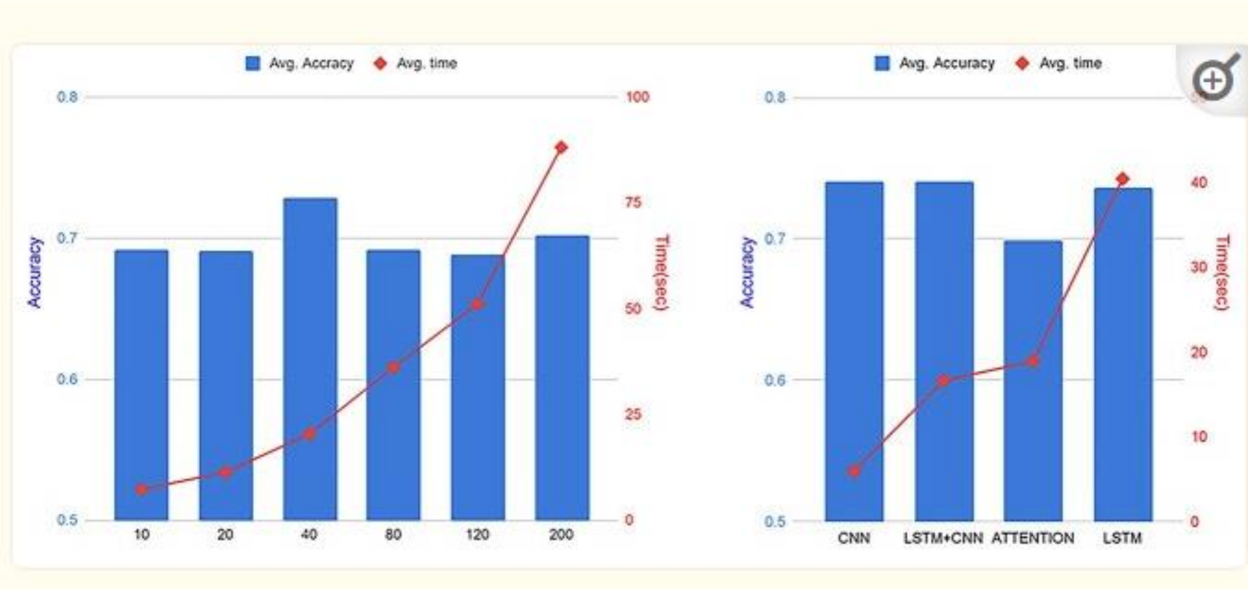

Fig 1. *Optimal fragment size and the most suitable deep learning algorithm were selected through the assessment of mean accuracy and computation time for Alzheimer's disease (AD) classification. This involved analyzing various fragment sizes, ranging from 10 to 200 SNPs within the APOE region, and employing different deep learning algorithms (CNN, LSTM, LSTM-CNN, and attention). The results presented in Figure (A) demonstrate the average accuracy and time in relation to the fragment size. Notably, the highest accuracy for AD classification was observed with a fragment containing 40 SNPs across CNN, LSTM-CNN, and LSTM models. Although the accuracy difference was minimal across different window sizes, processing time exhibited an increase with larger window sizes.*

Figure (B) showcases the average accuracy and time as functions of the deep learning algorithm, using a window size of 40. It is evident that the computation time for LSTM, LSTM-CNN, and attention models experienced a significant surge compared to CNN due to the inclusion of a greater number of SNPs in their respective fragments.

In the second phase, a Sliding Window Association Test (SWAT) was employed, and Convolutional Neural Network (CNN) was executed on the selected fragments. This process aimed to calculate the Phenotype Influence Score (PIS) of each Single Nucleotide Polymorphism (SNP) within the chosen fragments, identifying phenotype-associated SNPs based on their respective PIS values, as illustrated in Figure 2. For each SNP, a mean accuracy

Article DOI: 10.52589/AJMSS-4WNIT6F9
DOI URL: https://doi.org/10.52589/AJMSS-4WNIT6F9

of 40 windows, represented by the PIS of the SNP, was computed. Utilizing these PIS values, z-scores and one-tailed P-values were then calculated. The resulting Manhattan plot in Figure 3 displays the −log10 P-values on the y-axis against the SNP position in the genome on the x-axis. Notably, the SNP with the smallest P-value was identified as rs5117 in the APOC1 gene (P-value = 1.04E−22), followed by rs429358 in the APOE gene with a P-value of 1.41E-16. This genetic region encompassing APOE/APOC1/TOMM40 genes is widely recognized as the most robust genetic risk locus for Alzheimer's disease (AD) [30, 82–84]. Additionally, other significant genetic loci were identified at SNX14, SNX16, BICD1, WDR72, and GLT1D1 genes.

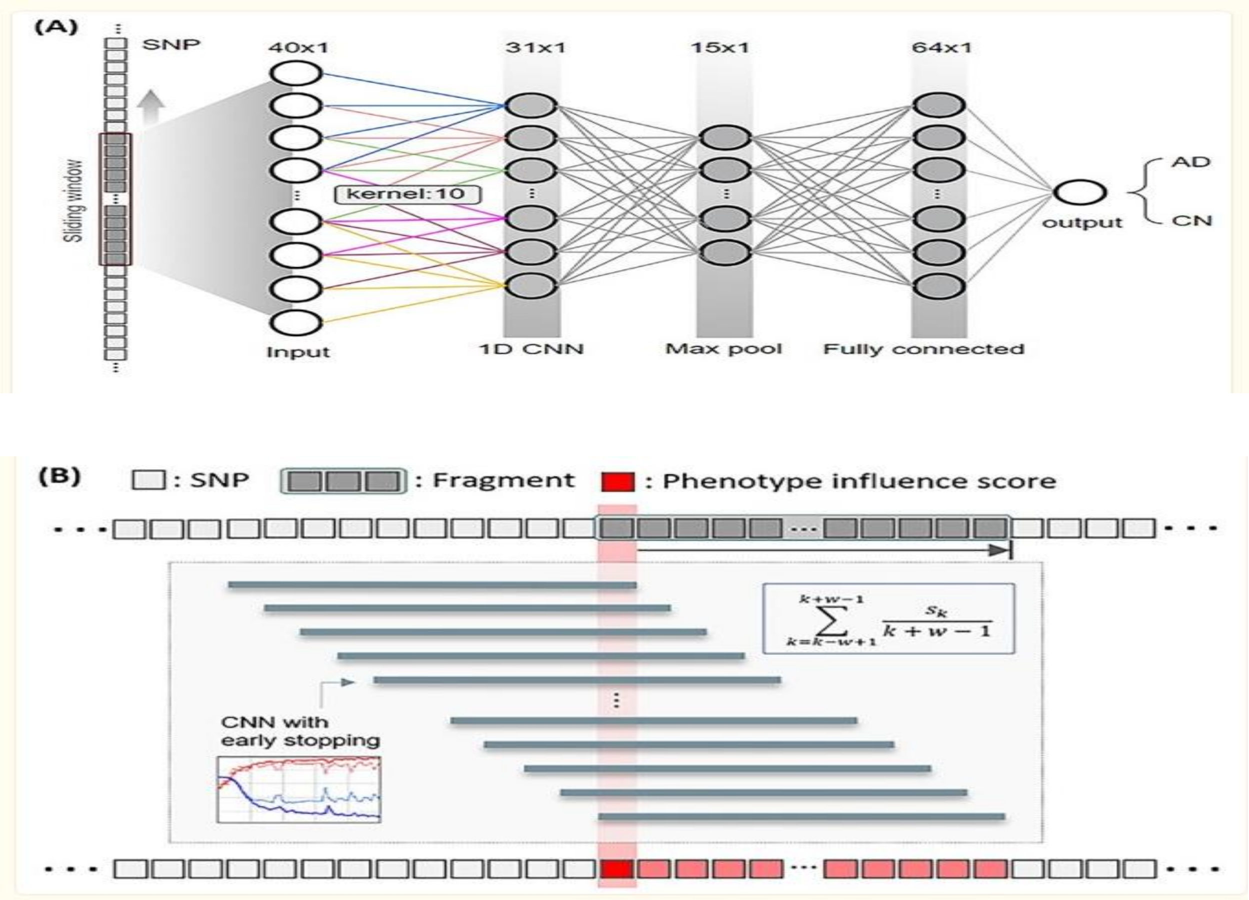

Fig 2. *Sliding Window Association Test (SWAT) for genetic variants. (A) Inside view of a sliding window that traverses the entire genome sequence to find a location that is associated with a specific phenotype. A CNN consisting of a convolutional layer with a kernel size of 10, a pooling layer with a maximum pool size of 2, a fully connected layer of 64 nodes, and an output layer with softmax activation was used. (B) 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 Sk is the kth SNP in the fragment.*

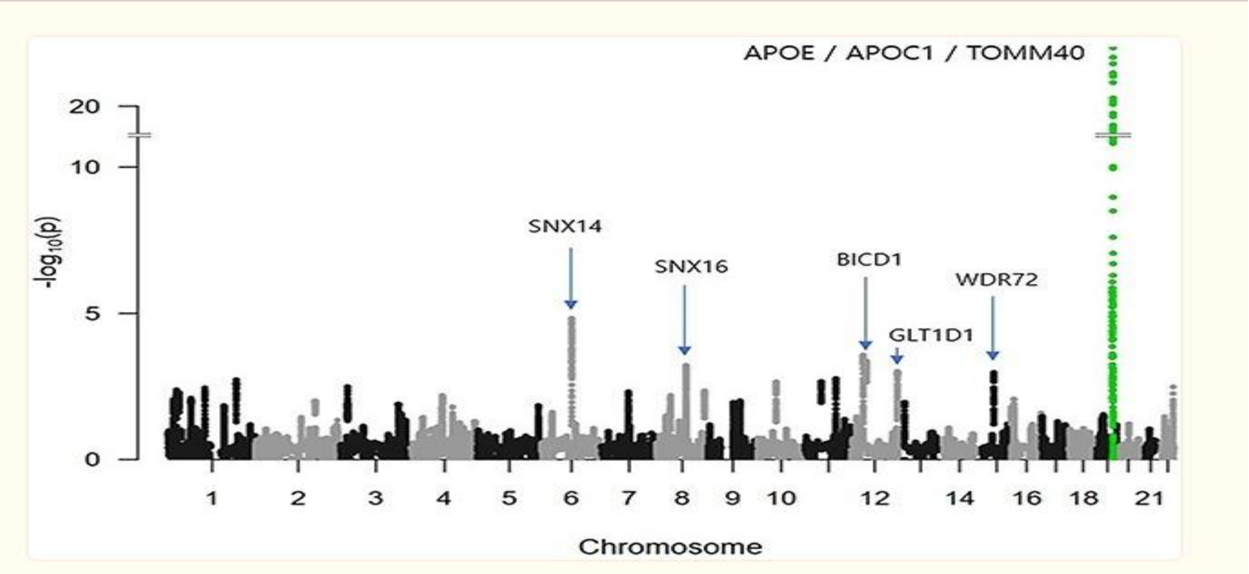

Fig 3. *A Manhattan plot is depicted, where the X-axis represents the positions of SNPs in the genome, and the Y-axis displays the -log10 of P-values. Notably, the genetic region encompassing APOE, APOC1, and TOMM40 genes is recognized as the most robust genetic risk locus for Alzheimer's disease. The SNP with the most significant P-value was identified as rs5117 in the APOC1 gene, with a P-value of 1.04E−22. Another noteworthy SNP, rs429358 in APOE, exhibited a P-value of 1.41E−16. Additionally, the subsequent genetic loci were identified at SNX14, SNX16, BICD1, WDR72, and GLT1D1 genes.*

In the third step, Convolutional Neural Network (CNN) was executed on the identified Single Nucleotide Polymorphisms (SNPs) to construct an Alzheimer's disease (AD) classification model. The classification outcomes of AD versus cognitively normal (CN) individuals are presented in Table 1, utilizing subsets ranging from the top 100 to 10,000 SNPs based on Phenotype Influence Scores (PIS). To facilitate comparison with conventional machine learning techniques, XGBoost and random forest were employed as classifiers.

The highest mean accuracy achieved through 10-fold cross-validation for AD classification by CNN was 75.02%, with an Area Under the Curve (AUC) of 0.8157, observed for a subset containing 4000 SNPs. This result indicated a 6.3% higher accuracy compared to random forest for a subset with 2000 SNPs and a 1.94% higher accuracy than XGBoost for a subset with 1000 SNPs. Notably, when classifying AD using only the count of APOE ε4 alleles, the accuracy was 66.7%, reflecting an 8.3% lower accuracy than our proposed method. In all instances, our CNN models demonstrated superior performance compared to the two traditional machine learning models, random forest and XGBoost, as illustrated in Figure 4.

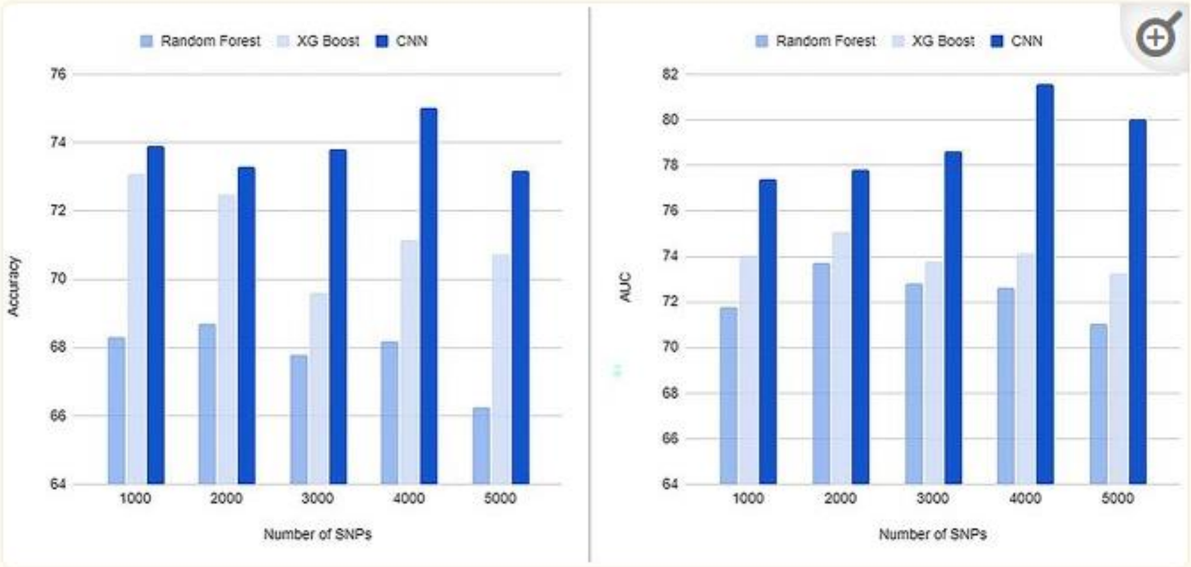

Fig 4. *Classification Results for AD versus CN. The horizontal axis depicts the quantity of top SNPs chosen using the phenotype influence score for AD classification. The vertical axis illustrates the accuracy (A) and AUC (B) derived from 10-fold cross-validation. Our CNN-based approach demonstrated the highest accuracy and AUC at 75.02% and 0.8157, respectively, with a selection of 4000 SNPs. In every instance, our CNN models surpassed the performance of two conventional machine learning models, random forest, and XGBoost.*

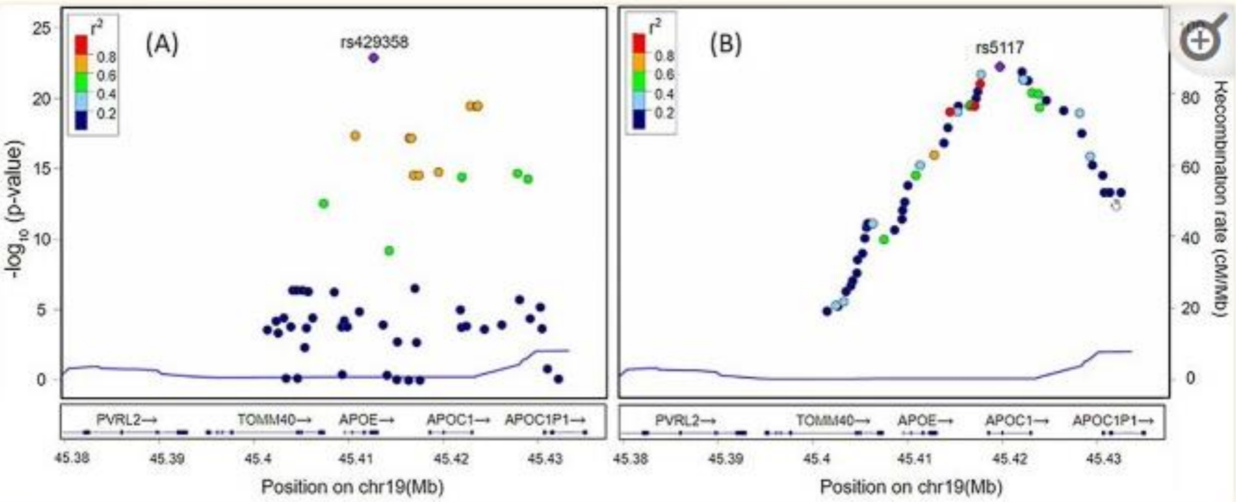

Fig 5. *LocusZoom plots depicting SNPs within the 300 kb upstream and downstream region from the boundary of the APOE gene. The horizontal axis represents SNP locations, and the vertical axis represents the -log10 of P-values. Each dot on the plot signifies an SNP, with the color indicating the squared correlation coefficient (r2) with the most significant SNP. (A) illustrates P-values computed using PLINK, with the most significant SNP identified as rs429358 in APOE. (B) displays P-values computed using our deep learning approach,*

Article DOI: 10.52589/AJMSS-4WNIT6F9
DOI URL: https://doi.org/10.52589/AJMSS-4WNIT6F9

identifying rs5117 in APOC1 as the most significant SNP. Notably, in (B), a linear increase is observed on the left side of rs5117, contrasting with a linear decrease on the right side of rs5117. This pattern differs from PLINK results in (A), which lack such linear trends. Additionally, in (B), three strongly correlated SNPs (r2 > 0.8) with rs5117 are evident on the left side, while no SNPs on the right side exhibit such correlations.

DISCUSSION

While deep learning has proven effective in addressing various real-world challenges, its application in Genome-Wide Association Studies (GWAS) and sequence data for genetic variant identification and disease/risk classification has been limited, primarily due to the high dimensionality of genomic data [22]. In this investigation, we introduce a novel deep learning-based sliding window approach designed to identify and select disease-associated Single Nucleotide Polymorphisms (SNPs) and construct an accurate classification model using high-dimensional genomic data, validated on the Alzheimer’s Disease Neuroimaging Initiative (ADNI) cohort (N = 981). Our method successfully pinpointed significant genetic loci for Alzheimer's disease (AD), including the well-established APOE genetic locus, and revealed additional risk loci. Notably, our deep learning-based approach demonstrated compatibility with traditional machine learning methods in AD classification.

The three-step deep learning-based approach for genetic variant identification begins with the division of the entire genome into non overlapping fragments of an optimal size, presenting an innovative fragmentation and windowing methodology, the first of its kind in deep learning-based genetic variant identification.

In the second step, we employed an overlapping window and Convolutional Neural Network (CNN) algorithm to compute a Phenotype Influence Score (PIS) for each SNP within the selected fragments. PIS serves as a novel index for identifying disease-related variants and predicting disease outcomes. Additionally, z-scores and one-tailed P-values were calculated using PIS, leading to a Manhattan plot showcasing the most significant genetic loci, particularly within the APOE/APOC1/TOMM40 genes, renowned as robust genetic risk factors for AD. Our method also brought to light several novel candidate genetic loci, such as sorting nexin (SNX) 14 and SNX16, situated on chromosomes 6 and 8, respectively, not previously associated with AD.

The third step involves the selection of top SNPs based on PIS to construct classification models for AD. Sets of highly AD-related SNPs were chosen, and CNN, along with traditional machine learning algorithms, XGBoost and random forest, were employed for classification. Classification accuracy varied with the number of selected SNPs and the classification algorithms. CNN exhibited the highest mean accuracy of 75.0% when applied to the top 4000 SNPs, comparable to traditional machine learning algorithms. Importantly, this accuracy surpassed that achieved by considering only the number of APOE ε4 alleles, indicating the efficacy of our proposed method. Notably, our method demonstrated the potential for extracting SNPs related to a phenotype by considering surrounding SNPs, contrasting with PLINK results that did not show SNPs with r2 greater than 0.8.

In conclusion, our innovative deep learning-based approach presents a robust means to identify AD-related SNPs and construct a classification model using genome-wide data. Given the estimated heritability of AD up to 80%, the identification of novel genetic loci related to the disease is crucial. Despite a modest sample size, our method identified a significant genetic locus with a classification accuracy of 75%. Future endeavors will involve applying our approach to large-scale whole genome sequencing datasets, refining classification models and exploring early stages of the disease for more nuanced risk assessment. Additionally, investigations into quantitative endophenotypes will contribute valuable insights into specific disease pathways and mechanisms.

ACKNOWLEDGEMENT

This project's data collection and sharing were made possible through the Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI receives support from the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, as well as generous contributions from various organizations.

CONCLUSION

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 (SWAT-CNN) for identification of genetic variants using deep learning to identify phenotype-related single nucleotide polymorphisms (SNPs) that can be applied to develop accurate disease 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.

REFERENCES

Alipanahi B, Delong A, Weirauch MT, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;33:831–8. [PubMed] [Google Scholar]

Amaldi E, Kann V. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theor Comput Sci* 1998;209:237–60. [Google Scholar]

Angermueller C, Lee HJ, Reik W, et al. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol* 2017;18:67. [PMC free article] [PubMed] [Google Scholar]

Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature* 2015;526:68–74. [PMC free article] [PubMed] [Google Scholar]

Article DOI: 10.52589/AJMSS-4WNIT6F9
DOI URL: https://doi.org/10.52589/AJMSS-4WNIT6F9

Bellomo G, Indaco A, Chiasserini D, et al. Machine learning driven profiling of cerebrospinal fluid Core biomarkers in Alzheimer's disease and other neurological disorders. *Front Neurosci* 2021;15:337. [PMC free article] [PubMed] [Google Scholar]

Bourdenx M, Martín-Segura A, Scrivo A, et al. Chaperone-mediated autophagy prevents collapse of the neuronal metastable proteome. *Cell* 2021;184:2696, e2625–714. [PMC free article] [PubMed] [Google Scholar]

Breiman L. Random forests. *Mach Learn* 2001;45:5–32. [Google Scholar]

Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 2019;47:D1005–12. [PMC free article] [PubMed] [Google Scholar]

Canter RG, Penney J, Tsai L-H. The road to restoring neural circuits for the treatment of Alzheimer's disease. *Nature* 2016;539:187–96. [PubMed] [Google Scholar]

Cervantes S, Samaranch L, Vidal-Taboada JM, et al. Genetic variation in APOE cluster region and Alzheimer's disease risk. *Neurobiol Aging* 2011;32:2107.e2107–17. [PubMed] [Google Scholar]

Chia R, Sabir MS, Bandres-Ciga S, et al. Genome sequencing analysis identifies new loci associated with Lewy body dementia and provides insights into its genetic architecture. *Nat Genet* 2021;53:294–303. [PMC free article] [PubMed] [Google Scholar]

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;261:921–3. [PubMed] [Google Scholar]

Ding Y, Sohn JH, Kawczynski MG, et al. A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain. *Radiology* 2019;290:456–64. [PMC free article] [PubMed] [Google Scholar]

Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. *J Mach Learn Res* 2011;12:2121–59. [Google Scholar]

Fang EF, Hou Y, Palikaras K, et al. Mitophagy inhibits amyloid-β and tau pathology and reverses cognitive deficits in models of Alzheimer's disease. *Nat Neurosci* 2019;22:401–12. [PMC free article] [PubMed] [Google Scholar]

Farrer LA, Cupples LA, Haines JL, et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: a meta-analysis. *JAMA* 1997;278:1349–56. [PubMed] [Google Scholar]

Felsky D, Roostaei T, Nho K, et al. Neuropathological correlates and genetic architecture of microglial activation in elderly human brain. *Nat Commun* 2019;10:1–12. [PMC free article] [PubMed] [Google Scholar]

Freedman ML, Reich D, Penney KL, et al. Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004;36:388–93. [PubMed] [Google Scholar]

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. *Proc Natl Acad Sci* 2014;111:E3604–13. [PMC free article] [PubMed] [Google Scholar]

Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*; 2011;15:315–23. [Google Scholar]

Goodfellow I, Bengio Y, Courville A, et al. *Deep Learning*. MIT press Cambridge, 2016. [Google Scholar]

Article DOI: 10.52589/AJMSS-4WNIT6F9
DOI URL: https://doi.org/10.52589/AJMSS-4WNIT6F9

Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res* 2003;3:1157–82. [Google Scholar]

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;9:1116–29. [PubMed] [Google Scholar]

Hinton G, Srivastava N, Swersky K. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. Coursera Lect Slides. 2012;14. [Google Scholar]

Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;9:1735–80. [PubMed] [Google Scholar]

Horgusluoglu E, Nudelman K, Nho K, et al. Adult neurogenesis and neurodegenerative diseases: a systems biology perspective. *Am J Med Genet B Neuropsychiatr Genet* 2017;174:93–112. [PMC free article] [PubMed] [Google Scholar]

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. [PMC free article] [PubMed] [Google Scholar]

Hyman BT, Phelps CH, Beach TG, et al. National Institute on Aging–Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease. *Alzheimers Dement* 2012;8:1–13. [PMC free article] [PubMed] [Google Scholar]

Jo T, Cheng J. Improving protein fold recognition by random forest. *BMC Bioinformatics* 2014;15:S14. [PMC free article] [PubMed] [Google Scholar]

Jo T, Nho K, Risacher SL, et al. Deep learning detection of informative features in tau PET for Alzheimer's disease classification. *BMC Bioinformatics* 2020;21:496. [PMC free article] [PubMed] [Google Scholar]

Jo T, Nho K, Saykin AJ. Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. *Front Aging Neurosci* 2019;11:220. [PMC free article] [PubMed] [Google Scholar]

Kerr JS, Adriaanse BA, Greig NH, et al. Mitophagy and Alzheimer's disease: cellular and molecular mechanisms. *Trends Neurosci* 2017;40:151–66. [PMC free article] [PubMed] [Google Scholar]

Kim HK, Min S, Song M, et al. Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity. *Nat Biotechnol* 2018;36:239–41. [PubMed] [Google Scholar]

Kingma DP, Adam BJ. Adam: A method for stochastic optimization. *arXiv preprint arXiv:14126980* 2014.

Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst* 2012;25:1097–105. [Google Scholar]

Lautrup S, Sinclair DA, Mattson MP, et al. NAD+ in brain aging and neurodegenerative disorders. *Cell Metab* 2019;30:630–55. [PMC free article] [PubMed] [Google Scholar]

LeCun Y, Touresky D, Hinton G, Sejnowski T. A theoretical framework for back-propagation. In: *Proceedings of the 1988 Connectionist Models Summer School*, CMU, Pittsburg, PA, 1988: 21–8.

Lee J, Retamal C, Cuitiño L, et al. Adaptor protein sorting nexin 17 regulates amyloid precursor protein trafficking and processing in the early endosomes. *J Biol Chem* 2008;283:11501–8. [PMC free article] [PubMed] [Google Scholar]

Article DOI: 10.52589/AJMSS-4WNIT6F9
DOI URL: https://doi.org/10.52589/AJMSS-4WNIT6F9

Lee Y-J, Han SB, Nam S-Y, et al. Inflammation and Alzheimer's disease. *Arch Pharm Res* 2010;33:1539–56. [PubMed] [Google Scholar]

Leenay RT, Aghazadeh A, Hiatt J, et al. Large dataset enables prediction of repair after CRISPR–Cas9 editing in primary T cells. *Nat Biotechnol* 2019;37:1034–7. [PMC free article] [PubMed] [Google Scholar]

Li F, Yang Y, Xing EP. From Lasso regression to feature vector machine. In: *Proceedings of the 18th International Conference on Neural Information Processing Systems*. 2005, pp. 779–86. MIT Press, Vancouver, British Columbia, Canada. [Google Scholar]

Listgarten J, Weinstein M, Kleinstiver BP, et al. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nat Biomed Eng* 2018;2:38–47. [PMC free article] [PubMed] [Google Scholar]

Liu J, Li L. Targeting autophagy for the treatment of Alzheimer's disease: challenges and opportunities. *Front Mol Neurosci* 2019;12:203. [PMC free article] [PubMed] [Google Scholar]

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 Comput Biol* 2019;15:e1007480. [PMC free article] [PubMed] [Google Scholar]

Mahoney ER, Dumitrescu L, Seto M, et al. Telomere length associations with cognition depend on Alzheimer's disease biomarkers. *Alzheimers Dement Transl Res Clin Interv* 2019;5:883–90. [PMC free article] [PubMed] [Google Scholar]

McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys* 1943;5:115–33. [PubMed] [Google Scholar]

Mercado N, Colley T, Baker JR, et al. Bicaudal D1 impairs autophagosome maturation in chronic obstructive pulmonary disease. *FASEB BioAdv* 2019;1:688–705. [PMC free article] [PubMed] [Google Scholar] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8921609/

Minsky M, Papert SA. *Perceptrons: An Introduction to Computational Geometry*. MIT press, 1969; [Google Scholar]

Morris JC, Roe CM, Xiong C, et al. APOE predicts amyloid-beta but not tau Alzheimer pathology in cognitively normal aging. *Ann Neurol* 2010;67:122–31. [PMC free article] [PubMed] [Google Scholar]

Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. *Icml* 2010:807–814. [Google Scholar]

Ogden PJ, Kelsic ED, Sinai S, et al. Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design. *Science* 2019;366:1139. [PMC free article] [PubMed] [Google Scholar]

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. [PMC free article] [PubMed] [Google Scholar]

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. [PMC free article] [PubMed] [Google Scholar]

Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 2010;26:2336–7. [PMC free article] [PubMed] [Google Scholar]

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;81:559–75. [PMC free article] [PubMed] [Google Scholar]

Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* 2016;44:e107–7. [PMC free article] [PubMed] [Google Scholar]

Rodriguez S, Hug C, Todorov P, et al. Machine learning identifies candidates for drug repurposing in Alzheimer's disease. *Nat Commun* 2021;12:1–13. [PMC free article] [PubMed] [Google Scholar]

Rosenblatt F. *The Perceptron, A Perceiving and Recognizing Automaton*. Technical Report 85–460-1, Cornell Aeronautical Laboratory, Buffalo, New York, 1957. [Google Scholar]

Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev* 1958;65:386. [PubMed] [Google Scholar]

Roses AD, Lutz MW, Amrine-Madsen H, et al. A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease. *Pharmacogenomics J* 2010;10:375–84. [PMC free article] [PubMed] [Google Scholar]

Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature* 1986;323:533–6. [Google Scholar]

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;43:1467–7. [PubMed] [Google Scholar]

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. [PMC free article] [PubMed] [Google Scholar]

Scherer M, Schmidt F, Lazareva O, et al. Machine learning for deciphering cell heterogeneity and gene regulation. *Nat Comput Sci* 2021;1:183–91. [Google Scholar]

Schwartzentruber J, Cooper S, Liu JZ, et al. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nat Genet* 2021;53:392–402. [PMC free article] [PubMed] [Google Scholar]

Sevigny J, Chiao P, Bussière T, et al. The antibody aducanumab reduces Aβ plaques in Alzheimer's disease. *Nature* 2016;537:50–6. [PubMed] [Google Scholar]

Shen MW, Arbab M, Hsu JY, et al. Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature* 2018;563:646–51. [PMC free article] [PubMed] [Google Scholar]

Sims R, Hill M, Williams J. The multiplex model of the genetics of Alzheimer's disease. *Nat Neurosci* 2020;23:311–22. [PubMed] [Google Scholar]

Stamate D, Kim M, Proitsi P, et al. A metabolite-based machine learning approach to diagnose Alzheimer-type dementia in blood: results from the European medical information framework for Alzheimer disease biomarker discovery cohort. *Alzheimers Dement Transl Res Clin Interv* 2019;5:933–8. [PMC free article] [PubMed] [Google Scholar]

Sutskever I, Martens J, Dahl G, Hinton G. On the importance of initialization and momentum in deep learning. In: *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013: 1139–47.

Suzanne M, Tong M. Brain metabolic dysfunction at the core of Alzheimer's disease. *Biochem Pharmacol* 2014;88:548–59. [PMC free article] [PubMed] [Google Scholar]

Article DOI: 10.52589/AJMSS-4WNIT6F9
DOI URL: https://doi.org/10.52589/AJMSS-4WNIT6F9

Tasaki S, Gaiteri C, Mostafavi S, et al. Deep learning decodes the principles of differential gene expression. *Nat Mach Intell* 2020;2:376–86. [PMC free article] [PubMed] [Google Scholar]

Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *arXiv preprint arXiv:1706.03762* 2017. [Google Scholar]

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;15:106–52. [PubMed] [Google Scholar]

Wainberg M, Merico D, Delong A, et al. Deep learning in biomedicine. *Nat Biotechnol* 2018;36:829–38. [PubMed] [Google Scholar]

Werbos PJ. Applications of advances in nonlinear sensitivity analysis. In: *System Modeling and Optimization*. Springer, Berlin, Heidelberg, 1982, vol 38, p 762–70. [Google Scholar]

Werbos PJ. Backwards differentiation in AD and neural nets: past links and new opportunities. In: *Automatic Differentiation: Applications, Theory, and Implementations*, Springer, Berlin, Heidelberg, 2006, vol 50, p 15–34.

Widrow B, Hoff ME. *Adaptive Switching Circuits*. Stanford Univ Ca Stanford Electronics Labs, 1960. WESCON Convention Record Part IV, 96–104. [Google Scholar]

Wong MW, Braidy N, Poljak A, et al. Dysregulation of lipids in Alzheimer's disease and their role as potential biomarkers. *Alzheimers Dement* 2017;13:810–27. [PubMed] [Google Scholar]

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. [PMC free article] [PubMed] [Google Scholar]

Xu Z, Huang G, Weinberger KQ, Zheng AX. Gradient boosted feature selection. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014, pp. 522–31. Association for Computing Machinery, New York, NY, USA. [Google Scholar]

Yamada M, Jitkrittum W, Sigal L, et al. High-dimensional feature selection by feature-wise Kernelized lasso. *Neural Comput* 2014;26:185–207. [PubMed] [Google Scholar]

Yan J, Qiu Y, Ribeiro dos Santos AM, et al. Systematic analysis of binding of transcription factors to noncoding variants. *Nature* 2021;591:147–51. [PMC free article] [PubMed] [Google Scholar]

Zhang J, Li Y, Tian J, Li T. LSTM-CNN hybrid model for text classification. In: *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC);*. 2018, 1675–80.

Zhang M, Schmitt-Ulms G, Sato C, et al. Drug repositioning for Alzheimer’s disease based on systematic ‘omics’ data mining. *PloS One* 2016;11:e0168812. [PMC free article] [PubMed] [Google Scholar]

Zhang S, Hu H, Jiang T, et al. TITER: predicting translation initiation sites by deep learning. *Bioinformatics* 2017;33:i234–42. [PMC free article] [PubMed] [Google Scholar]

Zhang Z, Park CY, Theesfeld CL, et al. An automated framework for efficiently designing deep convolutional neural networks in genomics. *Nature Machine Intelligence* 2021;3:392–400. [Google Scholar]

Zhao Y, Wang Y, Yang J, et al. Sorting nexin 12 interacts with BACE1 and regulates BACE1-mediated APP processing. *Mol Neurodegener* 2012;7:1–10. [PMC free article] [PubMed] [Google Scholar]

Zheng A, Lamkin M, Zhao H, et al. Deep neural networks identify sequence context features predictive of transcription factor binding. *Nat Mach Intell* 2021;3:172–80. [PMC free article] [PubMed] [Google Scholar]

Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods* 2015;12:931–4. [PMC free article] [PubMed] [Google Scholar]