IDENTIFICATION OF NEUROLOGICAL DISEASES USING DEEP LEARNING METHODS

Literature Review

Session: 2024-2024

Team:

Muhammad Tahir (21k-4503)
Insha Javed (21k-3279)
Hasan Iqbal (21k-3297)

Supervisor: Sir Shoaib Rauf

Department of Computer Science FAST-National University of Computer & Emerging Sciences, Karachi

Introduction:

The genetic basis for complex neurodegenerative diseases like Alzheimer's Disease (AD) has captivated researchers for decades. The emergence of genome-wide association studies (GWAS) has opened doors for the identification of single nucleotide polymorphisms (SNPs) related to disease susceptibility; however, despite the many candidate loci identified, predictive ability remains limited in traditional GWAS due to the complexity of genetic and environment factors, as well as the high-dimensionality of genomic data [1]. This work looks to improve SNP-based disease classification by using deep learning models including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, as well as a Phenotype Influence Score (PIS) for feature ranking.

This study has the potential to better interpret SNP-based classification models and enhance their performance. There is ample literature on GWAS and machine learning to support genomic analysis, however, there have been less studies that use end-to-end deep-learning models with feature attribution, particularly as it pertains to AD. This review includes peer-reviewed and conference papers that either utilize traditional statistical methods, classical machine learning models, or deep learning approaches to SNP selection or disease classification, as applied to AD or other complex diseases. Included studies were chosen for their relevance of methodology, newness of the technique, and application to genomic data. The review has five subsections: (1) traditional GWAS approaches, (2) Machine-learning methods for SNP classification, (3) Deep learning applications in genomics, (4) Deep learning feature selection techniques, and (5) limitations of existing research.

Traditional GWAS Approaches

Commonly used GWAS methods use tests such as logistic regression, chi-square tests and linear mixed models to locate SNPs associated with disease phenotypes [2]. Generally, these tests will consider each SNP as a separate component without exploring any interactions between SNPs, which limits the ability to detect a more complex pattern of association [3]. In AD in particular, while key loci such as the APOE allele have been implicated time and time again these methods have resulted in many SNPs with small effects remaining unidentified due to the use of univariate testing. Lastly, traditional GWAS methods are limited by the multiple testing problem often resulting in researchers using simple correction terms such as Bonferroni adjustment resulting in a loss of statistical power.

Machine Learning Techniques for SNP Selection and Disease Classification

Conventional GWAS typically apply statistical methods, such as logistic regression, chi-square tests, and linear mixed models, to analyze the relationship between SNPs and disease phenotype and identify SNPs associated with the disease phenotype [2]. These methods usually require an analysis of each SNP alone; they do not provide a systematic way to assess how SNPs may interact with one another and thus have relatively low sensitivity to detect more complex patterns of association [3]. For AD, there are some loci, such as the APOE locus that have exhibited consistent association under traditional GWAS methods; however, the overwhelming SNPs with small effects have gone unnoticed because of the univariate nature of the tests. Also, there are problems with multiple testing, whereby researchers often apply a strict approach for controlling the false discovery rate through correction of p-values, such as the Bonferroni methods; this correction will lead to a loss of power.

Deep Learning in Genomic Data Analysis

Deep learning (DL) has recently been recognized as a strong tool for modeling genomic data. Since DL models can learn hierarchies of features automatically from raw inputs [7], these models have been successfully employed to model genomic data. CNNs are able to learn spatial dependencies in the SNPs, while recurrent models such as LSTM may be able to learn long-range dependencies. For example, Tan et al. [8] showed that a CNN-based architecture performed better than traditional classifiers on separating AD from mild cognitive impairment (MCI). Furthermore, models with hybrid architectures such as CNN-LSTM have been proposed that capture local and global patterns on SNP sequence [9]. Although effective, DL models are still generally black boxes, therefore it is difficult to know why certain features influence the prediction.

Feature Selection with Deep Learning

There is still a lot to figure out about selection procedures for features in the analysis of genomics data. A few authors have targeted interpretability into their work with deep learning pipelines. For instance, AD-Syn-Net utilizes a subnet discovery process and targets SNPs that have the most vulnerability for co-mutation for AD [10]. Huang et al. [11] worked with transferring learned representations of SNPs over multiple datasets. There have also been feature-wise kernelized LASSO [12] and automated CNN architecture approaches [13] to create more interpretable reduced dimensionality models that are not losing predictive utility. Despite these additional efforts not many models measure contributions of each SNP to the outcome in an interpretable way or even measure contributions at all.

Limitations and Research Gap

While substantial progress has been made in using ML and DL for SNP-based classification, key limitations persist. Many models rely on fixed-length genomic windows, potentially missing disease-relevant fragments outside of predefined regions. Furthermore, interpretability is often sacrificed for accuracy, with few studies providing a quantifiable measure of SNP importance related to phenotype expression. This research addresses these gaps by proposing a deep learning framework that analyzes SNP fragments of varying sizes using CNN-LSTM models and introduces a novel metric, Phenotype Influence Score (PIS), to identify and rank SNPs contributing most significantly to disease classification.

Conclusion

The body of literature reviewed provides a discussion of the shift from classical statistical approaches to machine learning and deep learning approaches for genomic data analysis. Classical GWAS approaches remain a prominent foundation, but challenge any exploration of SNP interaction effects due to their limitation in detecting SNP interactions of a complex nature. Although machine learning models can improve feature (SNP) selection and classification performance, they can be limited to use when specific domain and area of studies preprocessing variables are ignored. Deep learning approaches can perform automatic feature learning, and typically produce greater accuracies, however interpretation is often lacking and can be challenging to discuss and convey the scientific meaning. This review has created a defined gap, models that incorporate classification performance while explaining and ranking SNP are lacking. Future research should promote the use of models that incorporate classification performance aligned to an interpretability framework, such as the one illustrated and developed in this study.

References

1. Wang X, Zhang Y, Liu H, et al. Predicting early Alzheimer's with blood biomarkers and clinical features. Alzheimers Res Ther 2024;16:19–27.

2. Wang Z, Li J, Chen X, et al. Wide and deep learning-based approaches for classification of Alzheimer's disease using genome-wide association studies. IEEE Trans Biomed Eng 2023;70:2546–55.

3. Zhang Q, Zhang L, Zhao Y, et al. AD-Syn-Net: systematic identification of Alzheimer's disease-associated mutation and co-mutation vulnerabilities via deep learning. Nat Commun 2023;14:567–78.
4. Huang Y, Zhou X, Yang H, et al. Transfer learning for classification of Alzheimer's disease based on genome-wide data. Bioinformatics 2023;39:2993–3001.

5. Lee H, Wong M, Kim C, et al. An Alzheimer's disease gene prediction method based on the ensemble of genome-wide association study summary statistics. Neurogenetics 2022;23:145–56.

6. Zhang H, Wang R, Zhao Z, et al. A machine learning method to identify genetic variants potentially associated with Alzheimer's disease. Front Genet 2021;12:735248. 12

7. Tan Y, Xu X, Liu X, et al. Use of deep-learning genomics to discriminate healthy individuals from those with Alzheimer's disease or mild cognitive impairment. J Alzheimers Dis 2021;82:1403–15.

8. Liu X, Yang H, Zhang L, et al. Early detection of Alzheimer's disease based on single nucleotide polymorphisms (SNPs) analysis and machine learning techniques. J Clin Neurosci 2020;77:222–30.

9. Lee J, Cho K, Park S, et al. A deep learning-based approach to detect neurodegenerative diseases. IEEE Access 2020;8:142005–15.

10. Zhang W, Zhao L, Zhang Q, et al. Early detection and characterization of Alzheimer's disease in clinical scenarios using Bioprofile concepts and k-means. Alzheimers Dement 2011;7:115–23.

11. Wainberg M, Merico D, Delong A, et al. Deep learning in biomedicine. Nat Biotechnol 2018;36:829–38

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

13. i'm HK, Min S, Song M, et al. Deep learning improves the prediction of CRISPR–Cpf1 guide RNA activity. Nat Biotechnol 2018;36: 239–41.

14. 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.

15. Zhang S, Hu H, Jiang T, et al. TITER: predicting translation initiation sites by deep learning. Bioinformatics 2017;33:i234–42.

16. Auton A, Abecasis GR, Altshuler DM, et al.A global reference for human genetic variation. Nature 2015;526:68–74. 13

17. HochreiterS,SchmidhuberJ.Longshort-termmemory. NeuralComput1997;9:1735–80.

18. LaudrupS, SinclairDA, MattsonMP, et al.NAD+in brain aging and neurodegenerative disorders.Cell Metab 2019;30:630–55.