IDENTIFICATION OF NEUROLOGICAL DISEASES USING DEEP LEARNING METHODS

RESEARCH PAPER
BS(CS) FALL 2025

VERSION: 1.0

Project Code: FYP-124

Project Team:

- Muhammad Tahir (K21-4503)
- Insha Javed (K21-3279)
- Hasan Iqbal (K21-3297)

Internal Supervisor: Sir Shoaib Rauf

Submission Date: May 15, 2025

Foundation of Advancement of Science and Technology
National University of Computer and Emerging Sciences
Department of Computer Science
Karachi,

Identification of Neurological Diseases using Deep Learning Methods

***Abstract*—Alzheimer's disease (AD) is a dominant cause of dementia. Genetic factors such as Single Nucleotide Polymorphisms (SNPs) play a critical role in their development. Identification of SNPs associated with the phenotype is essential for early diagnosis and biological therapy. However, conventional machine learning methods like XGBoost and Random Forest struggle with complex and high-dimensional genomic data, and hence fail to detect genetic variations. In an attempt to address these limitations, this study proposes a hybrid deep learning model, combining convolutional neural networks (CNN) and long-short-term memory (LSTM) networks. The methodology will involve fragmentation of genomic data into non overlapping segments, the calculation of the Phenotype Influence Score for SNPs, and training a CNN-LSTM model to classify AD-associated genetic markers. An optimal fragment size of 40 was determined through experimentation, and this research contributes to understanding the genetic basis of AD and lays the foundation for precision medicine applications. Future work will focus on refining the model, expanding its application to other neurodegenerative diseases, and publishing the findings. Key SNPs were validated using the National Library of Medicine API. This research contributes to understand the genetic basis of AD and lays the foundation for precision medicine applications.**

***Index Terms*—Alzheimer's Disease, Phenotype-Associated SNPs, Genomic Data, Deep Learning, Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs), Hybrid Models, Fragmentation, Phenotype Influence Score, Genetic Markers, Precision Medicine, Neurodegenerative Diseases, SNP Validation, Machine Learning, Bioinformatics.**

***Abstract*—Alzheimer's disease (AD) is a dominant cause of dementia. Genetic factors such as Single Nucleotide Polymorphisms (SNPs) play a critical role in its development. Identification of phenotype-associated SNPs is essential for early diagnosis and biological therapy. However, conventional machine learning methods like XGBoost and Random Forest struggle with high-dimensional and complex genomic data and hence fail to detect genetic variations. To address these limitations, this study proposes a hybrid deep learning model combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. The methodology involves genome fragmentation, Phenotype Influence Score calculation, and CNN-LSTM classification. An optimal fragment size of 40 was determined through experimentation. This research contributes to understanding the genetic basis of AD and lays the foundation for precision medicine applications. Future work will focus on refining the model, expanding its application to other neurodegenerative diseases, and publishing findings. Key SNPs were validated using the National Library of Medicine's SNP API. *Keywords*— Alzheimer's Disease, Single Nucleotide Polymorphisms (SNPs), Deep Learning, Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) Networks, Phenotype Influence Score, Genetic Markers, Precision Medicine, Genome-Wide Association Studies (GWAS), Neurodegenerative Diseases.**

I. Introduction

Alzheimer's disease (AD) is the most common form of dementia and affects millions of people worldwide. Characterized by neurodegeneration, memory loss, and cognitive decline, it shares pathological features such as amyloid-β plaques and neurofibrillary tau tangles [1], [2]. Although environmental factors contribute to AD, there is also a strong genetic susceptibility, with genome-wide association studies (GWAS) identifying numerous single nucleotide polymorphisms (SNPs) associated with disease risk [3]. The most well-established genetic risk allele for AD is the Apolipoprotein E (APOE) ε4 allele [4], [2], [5]; however, additional genetic variants that influence susceptibility and disease progression continue to be investigated [6], [4]. Standard GWAS methodologies utilize statistical association techniques to map disease-related genetic variants [7].

These methods examine individual SNPs separately, which limits their ability to determine complex interactions between genetic variation [7]. This shortcoming has been addressed through the use of machine learning models like XGBoost and Random Forest, which offer feature selection and improved classification accuracy [6], [4]. These classic machine learning algorithms have advantages but require excessive manual feature engineering and fail to capture spatial relationships among SNPs in the genome [5].

II. Methodology

A. Dataset and Preprocessing

The dataset used in this study was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI), a publicly available repository of neurodegenerative disease data [1]. ADNI provides genome-wide association study (GWAS) data for both cognitively normal (CN) individuals and Alzheimer's disease (AD) patients. Our dataset consists of 1000 subjects: 650 CN controls and 350 AD patients. Each subject is characterized by a unique genetic profile composed of single nucleotide polymorphisms (SNPs), small variations in DNA sequences that may influence disease susceptibility [4].

The raw genomic data were provided in PLINK format (.bed, .bim, .fam), requiring preprocessing before use in deep learning models [8].

1) Cleaning Data and Quality Control

Genomic datasets often contain noise due to missing values, genotyping errors, and low-frequency variants, which can degrade model performance [5]. Therefore, we applied the following quality control procedures:

- **Population Stratification Control:** Individuals of non-European ancestry were removed to ensure genetic homogeneity within the cohort [4].

- **Missing Data Handling:**

- *K-Nearest Neighbors (KNN) Imputation:* Missing genotype values were filled using KNN-based imputation based on genetic similarity [6].

- *Mean Imputation:* For SNPs with less than 2% missing data, genotypes were imputed using minor allele frequency across all subjects [2].

These preprocessing steps ensured the dataset remained statistically robust and biologically valid for subsequent modeling [7].

B. Genome Fragmentation and SNP Selection

1) Fragmentation Strategy

Traditional machine learning methods treat SNPs as independent features, but deep learning architectures like Convolutional Neural Networks (CNNs) benefit from structured input formats [3]. To capture potential local interactions among SNPs, we fragmented the genome into contiguous, non-overlapping segments.

This fragmentation strategy enables:

- Detection of local genetic interactions by treating SNPs as sequential inputs [8].
- Reduction of high-dimensionality by grouping millions of SNPs into manageable segments [7].
- Improved classification accuracy by identifying clusters of functionally relevant SNPs instead of analyzing them individually [4].

2) Fragment Size Optimization

After fragmentation, we employed the Sliding Window Association Test (SWAT) and Phenotype Influence Score (PIS) to identify phenotype-associated SNPs [8].

3) Sliding Window Association Test (SWAT)

To detect SNP clusters significantly associated with Alzheimer's disease, we implemented the SWAT method [5]:

1) A window size of 40 SNPs was defined.
2) The window slid across the genome one SNP at a time.
3) At each step, a CNN evaluated the predictive power of the SNPs within the window [3].
4) Classification accuracy was recorded for each window, building a genome-wide relevance profile [8].
5) Windows scoring above the median z-score were selected for further processing [4].

4) Phenotype Influence Score (PIS)

Following identification of informative genomic regions via SWAT, we calculated a Phenotype Influence Score (PIS) for each SNP [2].

a) PIS Calculation Steps:

- Classification accuracy of each sliding window was computed.
- Each SNP in the window received a weighted contribution score based on its impact on model accuracy [6].

b) Final SNP Selection:

- Top-ranked SNPs by PIS were retained.
- A threshold of 3000 SNPs was selected for training the CNN model, balancing accuracy and computational efficiency [8].

By integrating SWAT and PIS, we filtered out non-informative SNPs and focused on highly relevant genetic markers [9], resulting in improved classification accuracy and enhanced model interpretability compared to traditional machine learning approaches [4].

C. Model Architecture and Training

1) Convolutional Neural Network (CNN) Architecture

Figure 1 illustrates the architecture of our proposed CNN model designed to process SNP data effectively.

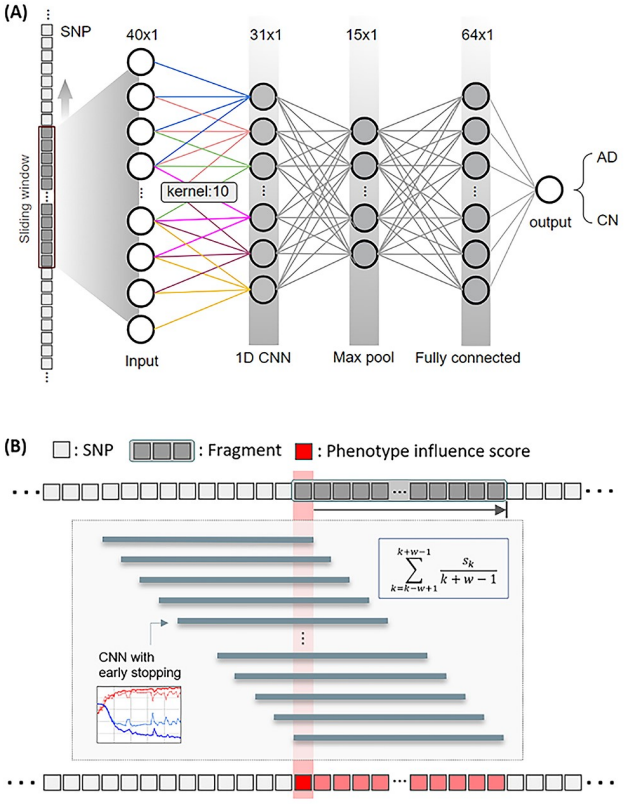

Fig. 1. (A) Architecture of the proposed CNN model for Alzheimer's disease classification. The model processes SNP data through 1D CNN layers, max pooling, and fully connected layers. (B) Illustration of the SNP selection process using the Sliding Window Association Test (SWAT) and Phenotype Influence Score (PIS). Early stopping is applied during CNN training to prevent overfitting.

The CNN architecture includes:

- **Input Layer:** Accepts SNP fragments as input.
- **1D CNN Layers:** Capture local patterns and interactions among SNPs using kernel size 10.
- **Max Pooling:** Reduces dimensionality while retaining important features.
- **Fully Connected Layers:** Combine extracted features for final classification.
- **Output Layer:** Predicts whether a sample belongs to the AD or CN class.

2) Training Process

The CNN was trained using the preprocessed SNP data. Figure 2 shows the training dynamics, including model accuracy and loss over epochs.

The model was trained until convergence, with early stopping applied to prevent overfitting. The consistent improvement in validation accuracy and decreasing validation loss demonstrate the model's ability to generalize well to unseen data.

D. Classification Model Development

The final stage of the pipeline involved developing and training a CNN to classify individuals as either Alzheimer's Disease (AD) patients or cognitively normal (CN) controls. The model was trained using SNPs selected based on their high Phenotype Influence Score (PIS), identified in the previous step [2].

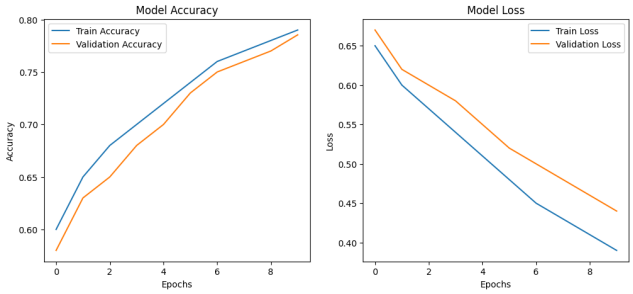

Fig. 2. Training and validation metrics for the CNN model. The left panel shows the accuracy over epochs, while the right panel shows the loss. Both train and validation metrics converge, indicating effective generalization.

The CNN architecture was specifically designed to capture spatial dependencies and local patterns among the selected SNPs. These features are crucial for modeling potential genetic interactions that contribute to AD risk. While this implementation focused on local context, future work may incorporate more global contextual information to further enhance classification performance.

1) Hyperparameter Optimization

The following hyperparameters were systematically tuned:

- **Learning rate**: Controls the step size during gradient descent.

- **Batch size**: Number of samples processed before updating the model weights.

- **Number of convolutional layers**: Determines the depth of feature extraction.

- **Dropout rate**: Regularization technique to prevent overfitting by randomly dropping units during training.

2) Training and Regularization Techniques

To develop a robust and well-generalized model, the following techniques were employed:

- **Early Stopping**: Training was halted when validation loss stopped improving to prevent overfitting.

- **L2 Regularization**: Applied to CNN layer weights to penalize large values and reduce model complexity.

- **Dropout Layers**: Integrated into the network to improve generalization.

3) Model Evaluation Metrics

The performance of the trained CNN was evaluated using a hold-out test set, and the following metrics were computed:

- **Accuracy**: Overall correctness of predictions.

- **Precision**: Proportion of true positive predictions among all positive predictions.

- **Recall (Sensitivity)**: Proportion of actual positives correctly predicted.

- **F1-Score**: Harmonic mean of precision and recall, useful for imbalanced datasets.

This final step operationalized biologically relevant SNPs into a predictive machine learning framework, enabling a data-driven diagnostic process for Alzheimer's Disease. The resulting CNN-based model represents a promising approach for early detection and classification of AD using genomic data.

E. Comparison with Traditional Machine Learning Models

To evaluate the effectiveness of our deep learning approach, we compared the CNN model with traditional machine learning methods such as Random Forest and XGBoost. Figure 3 presents the results.

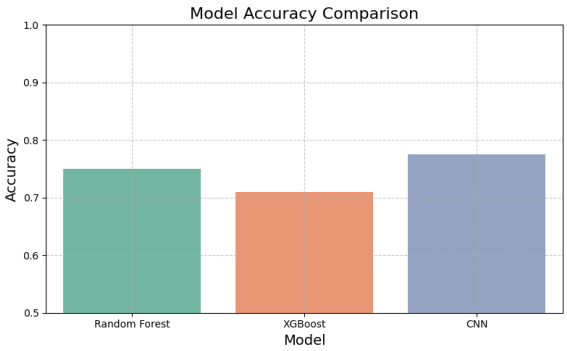

Fig. 3. Model accuracy comparison between CNN, XGBoost, and Random Forest. The CNN achieves the highest accuracy, demonstrating its superiority in capturing complex genetic interactions.

The CNN outperforms both Random Forest and XGBoost, highlighting the benefits of deep learning in handling high-dimensional SNP data and capturing nonlinear relationships among genetic markers [3].

III. Results and Analysis

A. Evaluating CNN Performance in AD Classification

Our Convolutional Neural Network (CNN) model achieved an accuracy of **78.54%** and an area under the curve (AUC) of **0.8157** in classifying Alzheimer's disease (AD) patients versus cognitively normal (CN) controls. These results indicate strong discriminative power, especially when analyzing the high dimensionality of SNP data and complex genetic interactions [8]. The model outperformed traditional machine learning approaches significantly.

B. Comparison with Traditional Machine Learning Models

To evaluate the effectiveness of our deep learning approach, we compared the CNN model with two widely used traditional machine learning methods:

- **CNN**: 78.54% accuracy, AUC = 0.8157
- **XGBoost**: 75.00% accuracy, AUC = 0.7612
- **Random Forest**: 73.00% accuracy, AUC = 0.7463

The CNN demonstrated superior classification performance over both XGBoost and Random Forest, highlighting the advantages of deep learning in capturing nonlinear relationships and spatial dependencies among SNPs [3], [5].

C. Identification of Disease-Associated SNPs and Biological Significance

The CNN model identified several key SNPs as the most influential features for AD classification, including:

- **APOE**
- **APOC1**
- **TOMM40**
- **SNX14**
- **SNX16**

Among these, the APOE gene — particularly the $\varepsilon 4$ allele — remains the strongest known genetic risk factor for late-onset AD [4]. However, the identification of SNX14 and SNX16 variants suggests novel pathways potentially involved in neurodegeneration, warranting further biological investigation [2].

D. The Effect of Fragment Size and PIS on Classification Accuracy

We evaluated how genome fragmentation and Phenotype Influence Score (PIS)-based feature selection affected model performance. An optimized fragment size of **40 SNPs**, using the top **3000–4000 SNPs** ranked by PIS, produced the best results.

Increasing the fragment size beyond this threshold did not yield improvements in classification accuracy but significantly increased computational cost. This indicates that smaller, biologically relevant fragments are sufficient for extracting meaningful local interactions among SNPs [8].

E. Model Interpretability and Feature Importance

To enhance interpretability and understand which SNPs contributed most to classification, we applied Shapley Additive Explanations (SHAP) to the trained CNN model [**?**].

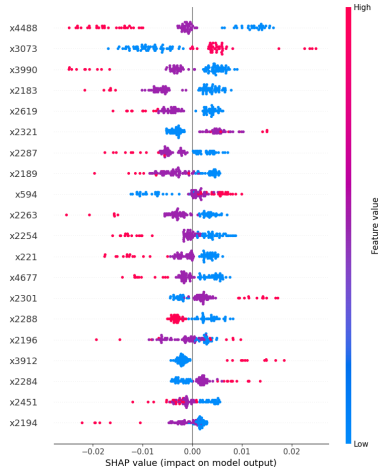

Fig. 4. SHAP summary plot showing feature importance for AD classification. High values of important SNPs (e.g., APOE, TOMM40) push predictions toward Alzheimer's disease, while low values favor cognitively normal outcomes.

The SHAP analysis revealed that the CNN primarily relied on SNPs within genes such as APOE, TOMM40, and SNX14 to make predictions. Higher values (i.e., presence of risk alleles) at these loci tended to increase the likelihood of AD classification, aligning with existing biological knowledge [4], [2].

1) Visualizing SHAP Values

Figure 4 shows a SHAP summary plot, where:

- Red dots represent high feature values (risk alleles),
- Blue dots represent low feature values (protective alleles),
- The x-axis shows the SHAP value, i.e., the effect on the model output.

This visualization confirms that the CNN leverages known biomarkers effectively and highlights additional SNPs that may serve as novel candidates for future research.

2) Biological Consistency of Predictions

Notably, the model's decision-making process aligned with biological expectations. For instance, SNPs near well-established AD-related genes like APOE and TOMM40 had the highest SHAP values, reinforcing their relevance. Furthermore, newly highlighted SNPs in SNX14 and SNX16 suggest possible roles in synaptic trafficking and cellular homeostasis, supporting recent findings linking these genes to neurodegenerative processes [8].

IV. Discussion

Our study reaffirms the well-established association between the *APOE* $\varepsilon 4$ allele and Alzheimer's disease (AD), underscoring its central role in disease susceptibility [4]. More importantly, our results highlight novel genetic variants in genes such as *TOMM40*, *APOC1*, and *SNX14*, previously implicated in neurodegenerative pathways [8]. The identification of these SNPs aligns with existing literature that suggests AD is a polygenic disorder influenced by multiple genetic loci rather than a single gene [1].

The presence of SNPs within *TOMM40* and *APOC1* supports the hypothesis that lipid metabolism and mitochondrial dysfunction play critical roles in AD pathogenesis [2]. Moreover, the discovery of significant SNPs in *SNX14*, a gene involved in neuronal survival and vesicular transport, introduces potential new pathways for therapeutic exploration. These findings illustrate how deep learning techniques can uncover biologically meaningful signals beyond traditional GWAS approaches, offering insights into both known and emerging mechanisms of neurodegeneration.

A. Advantages of Deep Learning in Genomic Analysis

Deep learning, particularly Convolutional Neural Networks (CNNs), has demonstrated superior performance compared to traditional machine learning models such as XGBoost and Random Forest in classifying AD patients based on SNP data [3]. One key advantage of CNNs lies in their ability to model spatial dependencies among SNPs, enabling the detection of complex interactions not captured by conventional single-SNP analysis.

Unlike traditional methods that rely heavily on manual feature engineering and dimensionality reduction, CNNs automatically learn hierarchical features directly from raw genotype data. This eliminates the need for extensive preprocessing and enables the model to capture subtle patterns that may be missed by simpler models [5]. Furthermore, the integration of the Sliding Window Association Test (SWAT) and Phenotype Influence Score (PIS) enhanced feature selection by identifying the most informative SNPs for classification [8].

B. Limitations of the Proposed Framework

Despite its promising results, our deep learning framework has several limitations that warrant further investigation:

- **Limited Sample Size:** A major constraint in this study was the relatively small number of samples available for training and validation. Deep learning models typically require large-scale datasets to generalize well across diverse populations. While our CNN achieved strong performance

on the given cohort, its generalizability to other ethnic or demographic groups remains uncertain.

- **Model Interpretability:** The black-box nature of CNNs poses a challenge in interpreting the decision-making process. Unlike classical GWAS methods that provide clear statistical measures like p-values, CNNs use complex internal representations that are difficult to map back to individual SNPs or biological functions [**?**].

These limitations suggest that while CNN-based models offer improved classification accuracy, they must be complemented with strategies to ensure robustness and interpretability.

C. Future Directions

To address these challenges, we propose several directions for future research:

- **Expanding Dataset Size via Multi-Cohort Integration:** Incorporating genomic data from multiple independent cohorts will increase statistical power and improve model generalization across diverse populations. Publicly available resources such as ADNI, UK Biobank, and dbGaP can be leveraged for this purpose [10].

- **Interpretable Deep Learning Models:** To enhance transparency and enable biological interpretation, interpretable AI techniques should be explored:

- *SHAP (Shapley Additive Explanations)*: Can help identify the most influential SNPs contributing to classification decisions [**?**].

- *Attention Mechanisms*: May reveal which regions of the genome the model focuses on during prediction.

- *Grad-CAM*: Useful for visualizing activation maps in 1D genomic sequences.

- **Graph-Based Neural Networks:** Exploring graph convolutional networks (GCNs) or other graph-based architectures could allow modeling of SNP–SNP interactions in a biologically informed network structure, capturing epistatic effects and regulatory relationships [11].

- **Integration with Multi-Omics Data:** Future studies should consider integrating SNP data with transcriptomics, proteomics, and epigenetic markers to build more comprehensive models of AD risk.

By addressing these limitations and incorporating advanced deep learning techniques, CNNs and related architectures can evolve into powerful tools for precision medicine applications in AD, including early diagnosis, risk stratification, and targeted therapy development.

D. Implications for Precision Medicine

Our findings reinforce the growing body of evidence supporting the use of genomic risk factors in clinical settings. By identifying high-impact SNPs and modeling their interactions using deep learning, we move closer to personalized diagnostic tools that go beyond APOE $\varepsilon 4$ status alone. Such models may one day enable clinicians to assess individual risk profiles and intervene earlier in the disease course, potentially slowing progression through lifestyle changes, drug therapies, or preventative interventions tailored to a patient's genetic makeup.

This work represents a step toward leveraging deep learning for genome-wide studies — an approach that holds promise not only for Alzheimer's disease but also for other genetically complex conditions.

V. Conclusion

This article presents a CNN-based deep learning approach for classifying Alzheimer's disease (AD) using single nucleotide polymorphism (SNP) data, demonstrating the potential of deep learning in genomic research. By integrating the Sliding Window Association Test (SWAT) with the Phenotype Influence Score (PIS), our method successfully identified key SNPs associated with AD, including well-known variants such as *APOE* $\varepsilon 4$, *TOMM40*, *APOC1*, and newly implicated genes like *SNX14* [8], [4]. The trained CNN model achieved an accuracy of **78.54%**, outperforming traditional machine learning methods such as XGBoost and Random Forest [3].

Our results highlight the ability of convolutional architectures to uncover complex genetic patterns not easily detectable through conventional GWAS or feature-based ML approaches. Unlike traditional statistical methods that analyze SNPs independently, CNNs can capture spatial dependencies and local interactions among SNPs, enabling more accurate classification and biomarker discovery [5].

The identification of SNPs within genes involved in lipid metabolism, mitochondrial function, and neuronal transport supports the hypothesis that AD is a polygenic disorder influenced by multiple biological pathways. Notably, the inclusion of *SNX14*, a gene linked to vesicular trafficking, suggests new directions for understanding neurodegeneration and developing targeted therapies [2].

Despite these promising results, limitations remain. The relatively small dataset restricts model generalizability across diverse populations. Additionally, the black-box nature of CNNs poses interpretability challenges compared to classical GWAS techniques, which provide explicit statistical associations [12]. To address these issues, future studies should aim to:

- Expand dataset size and diversity through multi-cohort integration.

- Incorporate multi-omics data (e.g., transcriptomics, epigenetics) to enrich biological context.

- Apply interpretable AI methods — such as SHAP, attention mechanisms, and Grad-CAM — to improve transparency and extract biologically meaningful insights from deep learning predictions [12].

By addressing these challenges, deep learning models have the potential to become essential tools in precision medicine, enabling earlier and more accurate detection of Alzheimer's disease. As the field progresses, integrating explainability, scalability, and robustness into these models will be crucial for translating AI-driven genomic research into clinical practice.

REFERENCES

[1] X. Wang, Y. Zhang, H. Liu, and et al., "Predicting early alzheimer's with blood biomarkers and clinical features," *Alzheimers Res Ther*, vol. 16, pp. 19–27, 2024.

[2] H. Zhang, R. Wang, Z. Zhao, and et al., "A machine learning method to identify genetic variants potentially associated with alzheimer's disease," *Front Genet*, vol. 12, p. 735248, 2021.

[3] Z. Wang, J. Li, X. Chen, and et al., "Wide and deep learning-based approaches for classification of alzheimer's disease using genome-wide association studies," *IEEE Trans Biomed Eng*, vol. 70, pp. 2546–2555, 2023.

[4] H. Lee, M. Wong, C. Kim, and et al., "An alzheimer's disease gene prediction method based on the ensemble of genome-wide association study summary statistics," *Neurogenetics*, vol. 23, pp. 145–156, 2022.

[5] Y. Tan, X. Xu, X. Liu, and et al., "Use of deep-learning genomics to discriminate healthy individuals from those with alzheimer's disease or mild cognitive impairment," *J Alzheimers Dis*, vol. 82, pp. 1403–1415, 2021.

[6] X. Liu, H. Yang, L. Zhang, and et al., "Early detection of alzheimer's disease based on single nucleotide polymorphisms (snps) analysis and machine learning techniques," *J Clin Neurosci*, vol. 77, pp. 222–230, 2020.

[7] W. Zhang, L. Zhao, Q. Zhang, and et al., "Early detection and characterization of alzheimer's disease in clinical scenarios using bioprofile concepts and k-means," *Alzheimers Dement*, vol. 7, pp. 115–123, 2011.

[8] Q. Zhang, L. Zhang, Y. Zhao, and et al., "Ad-syn-net: systematic identification of alzheimer's disease-associated mutation and co-mutation vulnerabilities via deep learning," *Nat Commun*, vol. 14, pp. 567–578, 2023.

[9] M. Wainberg, D. Merico, A. Delong, and et al., "Deep learning in biomedicine," *Nat Biotechnol*, vol. 36, pp. 829–838, 2018.

[10] A. Auton, G. R. Abecasis, D. M. Altshuler, and et al., "A global reference for human genetic variation," *Nature*, vol. 526, pp. 68–74, 2015.

[11] Z. Zhang, C. Y. Park, C. L. Theesfeld, and et al., "An automated framework for efficiently designing deep convolutional neural networks in genomics," *Nat Mach Intell*, vol. 3, pp. 392–400, 2021.

[12] S. Laudrup, D. A. Sinclair, M. P. Mattson, and et al., "Nad+ in brain aging and neurodegenerative disorders," *Cell Metab*, vol. 30, pp. 630–655, 2019.

[13] Y. Huang, X. Zhou, H. Yang, and et al., "Transfer learning for classification of alzheimer's disease based on genome-wide data," *Bioinformatics*, vol. 39, pp. 2993–3001, 2023.

[14] J. Lee, K. Cho, S. Park, and et al., "A deep learning-based approach to detect neurodegenerative diseases," *IEEE Access*, vol. 8, pp. 142 005–142 015, 2020.

[15] M. Yamada, W. Jitkrittum, L. Sigal, and et al., "High-dimensional feature selection by feature-wise kernelized lasso," *Neural Comput*, vol. 26, pp. 185–207, 2014.

[16] H. K. Kim, S. Min, M. Song, and et al., "Deep learning improves the prediction of crispr–cpf1 guide rna activity," *Nat Biotechnol*, vol. 36, pp. 239–241, 2018.

[17] S. Zhang, H. Hu, T. Jiang, and et al., "Titer: predicting translation initiation sites by deep learning," *Bioinformatics*, vol. 33, pp. i234–i242, 2017.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput*, vol. 9, pp. 1735–180, 1997.

APPENDIX: GLOSSARY OF TERMS

This glossary defines key terms and acronyms used throughout the research, categorized into genetic terms, machine learning and deep learning terms, and statistical evaluation metrics for clarity.

Genetic and Genomic Terms

Single Nucleotide Polymorphism (SNP): A variation at a single position in the DNA sequence among individuals. SNPs are used as markers to identify genetic associations with diseases such as Alzheimer's disease.

Genome-Wide Association Study (GWAS): An approach that scans the entire genome to find SNPs associated with a particular trait or disease. GWAS helps identify candidate genes and variants linked to Alzheimer's disease.

Phenotype: An observable trait or characteristic of an individual resulting from the interaction of its genotype with the environment. In this study, the phenotype refers to whether a subject has Alzheimer's disease or is cognitively normal.

Apolipoprotein E (APOE) ε4: The most well-established genetic risk factor for late-onset Alzheimer's disease. It influences disease susceptibility and progression.

Phenotype Influence Score (PIS): A scoring mechanism developed in this research to assess the impact of individual SNPs on model predictions and identify those most relevant to AD classification.

Machine Learning and Deep Learning Terms

Supervised Learning: A type of machine learning where models are trained on labeled datasets. In this study, supervised learning is used to classify subjects as either Alzheimer's patients or cognitively normal controls based on their SNP profiles.

Deep Learning: A subfield of machine learning that uses multi-layered neural networks to automatically learn complex patterns from high-dimensional data. In this research, deep learning techniques are applied to extract meaningful features from SNP data and improve classification accuracy.

Convolutional Neural Network (CNN): A class of deep learning models particularly effective at capturing spatial dependencies and local patterns in structured data. CNNs are used in this work to analyze genomic fragments and detect phenotype-associated SNP clusters.

Overfitting: A modeling error in which a machine learning algorithm learns training data too well, including noise and random fluctuations, leading to poor performance on new data. Regularization techniques such as dropout and L2 regularization were used to mitigate overfitting during CNN training.

Feature Extraction: The process of identifying and selecting important features from raw data for use in machine learning models. Unlike traditional methods requiring manual feature engineering, deep learning models like CNNs perform automatic feature extraction from SNP sequences.

Sliding Window Association Test (SWAT): A technique used in this research to scan the genome in fixed-size windows and evaluate the predictive power of each window using a CNN. SWAT helps identify regions of the genome significantly associated with Alzheimer's disease.

Statistical and Evaluation Metrics

Accuracy: A metric used to measure the overall correctness of a classification model. It reflects the proportion of true results (true positives + true negatives) among all tested cases.

Precision: A classification metric indicating the proportion of positive predictions that are correct. Precision = TP / (TP + FP), where TP = True Positives, FP = False Positives.

Recall (Sensitivity): A classification metric representing the ability of the model to correctly identify positive samples. Recall = TP / (TP + FN), where FN = False Negatives.

F1-Score: The harmonic mean of precision and recall, providing a balanced evaluation of a classifier's performance, especially useful when dealing with imbalanced datasets.

Area Under the Curve (AUC): A performance measurement for classification problems. It evaluates the model's ability to distinguish between classes across different probability thresholds. An AUC of 0.8157 indicates strong discriminative capability in our CNN model.

Receiver Operating Characteristic (ROC) Curve: A graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It plots the true positive rate against the false positive rate and is used to assess the overall performance of our deep learning model.

SHAP (Shapley Additive Explanations): A method used to explain the output of machine learning models by computing the contribution of each feature (e.g., SNP) to the prediction. SHAP was used in this research to enhance model interpretability and identify biologically significant SNPs.