FOUNDATION FOR ADVANCEMENT OF SCIENCE & TECHNOLOGY

NATIONAL UNIVERSITY OF COMPUTER & EMERGING SCIENCES

IDENTIFICATION OF NEUROLOGICAL DISEASES USING DEEP LEARNING METHODS

FINAL YEAR PROJECT REPORT

Supervisor: Sir Shoaib Rauf

Project Team:

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

Project Code: FYP-124

Batch: 2021

Department: Department of Computer Science

School: FAST School of Computing

Campus: Karachi, Sindh, Pakistan

Submission Date: May 15, 2025

Note: Submitted in fulfillment of the requirements for the degree of Bachelor of Computer Science.

Certificate of Approval

This is to certify that the project report entitled *"IDENTIFICATION OF NEUROLOGICAL DISEASES USING DEEP LEARNING METHODS"*, submitted by **Muhammad Tahir (K21-4503)**, **Insha Javed (K21-3279)**, and **Hasan Iqbal (K21-3297)** in partial fulfillment of the requirements for the degree of **Bachelor of Science in Computer Science**, has been examined and is hereby approved.

Sign-off Authority	Signature	Project Role	Sign-off Date
Sir Shoaib Rauf		Supervisor	May 15, 2025
Dr. Ghufran Ahmed		Head of Department	May 15, 2025

Declaration of Originality

We hereby declare that this project report titled *"IDENTIFICATION OF NEUROLOGICAL DISEASES USING DEEP LEARNING METHODS"* is our own original work. All sources of information and data have been acknowledged in full. This report has not been submitted previously, in whole or in part, for any other degree or qualification at this or any other institution. We attest that we have followed the university's academic integrity guidelines and that no part of this work is the result of plagiarism or unauthorized collaboration.

Team Members	Signature	Sign-off Date
Muhammad Tahir		May 15, 2025
Insha Javed		May 15, 2025
Hasan Iqbal		May 15, 2025

Document Information

Category	Information
Customer	National University of Computer and Emerging Sciences
Project	IDENTIFICATION OF NEUROLOGICAL DISEASES USING DEEP LEARNING METHODS
Document	Final Year Project Report
Status	Final
Author(s)	Muhammad Tahir, Insha Javed, Hasan Iqbal
Approver(s)	Sir Shoaib Rauf
Issue Date	May 15, 2025
Document Location	National University of Computer and Emerging Sciences

Distribution List

Name	Role
Sir Shoaib Rauf	Supervisor
Mr. Saad Manzoor	Project Coordinator

Acknowledgements

The successful completion of research on IDENTIFICATION OF NEUROLOGICAL DISEASES USING DEEP LEARNING METHODS would not have been possible without the unwavering support, guidance, and collaboration of numerous individuals and institutions. We extend our deepest gratitude to all who contributed to this journey.

To Our Project Team:
First and foremost, we thank our dedicated team members—Muhammad Tahir, Insha Javed and Hasan Iqbal—for their relentless effort, creativity, and perseverance. From late-night coding sessions to rigorous dataset validation, your commitment to excellence transformed this vision into reality.

To Our Mentor:
We are profoundly indebted to our Supervisor, Sir Shoaib Rauf, for his expert guidance, insightful critiques, and unwavering patience. His expertise in bioinformatics and deep learning provided the intellectual foundation for our work, while his encouragement motivated us to push boundaries even in the face of challenges.

To the Department and Institution:
We acknowledge the Department of Computer Science at the National University of Computer and Emerging Sciences (FAST-NUCES), Karachi, for providing state-of-the-art laboratories, computational resources, and an environment conducive to innovation. Special thanks to the faculty members who fostered our growth as researchers.

To Data Contributors:
We sincerely thank ADANI (Alzheimer's Disease Neuroimaging Initiative) for providing the dataset used in this research. Their support played a crucial role in enabling and advancing the findings of this study.

To Friends and Family:
Finally, we thank our families and friends for their unwavering emotional support. Your belief in our capabilities during moments of doubt kept us anchored, reminding us that every challenge was a stepping stone to success.

This project is a testament to collaboration, and we are humbled to have worked alongside such inspiring individuals.

Contents

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.**

1 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 [2, 4, 5]; however, additional genetic variants that influence susceptibility and disease progression continue to be investigated [4, 6]. 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 [4, 6]. 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].

2 Related Work

2.1 Genetic Association Studies and Machine Learning in Alzheimer's Disease

Numerous studies have focused on identifying genetic markers linked to Alzheimer's Disease (AD) using genome-wide association studies (GWAS). Research has highlighted the importance of SNPs in the *APOE* and *APOC1* regions, emphasizing their strong association with AD susceptibility [2, 4]. Traditional methods such as logistic regression and statistical analysis have been widely used to evaluate SNP significance; however, these approaches often lack the precision required to model complex genomic interactions [7].

Recently, machine learning techniques have been employed to enhance SNP detection. Random Forests and Support Vector Machines (SVMs) have shown promise in feature selection and classification tasks [4, 6]. While effective for certain applications, these models rely heavily on manual feature engineering and struggle to capture spatial or sequential dependencies among SNPs [5].

2.2 Deep Learning Approaches in Genomic Analysis

Deep learning methods, particularly Convolutional Neural Networks (CNNs), have emerged as powerful tools for analyzing structured genomic data due to their ability to automatically extract hierarchical features and detect local patterns [3]. CNNs are especially suited for modeling SNP clusters and capturing genetic interactions that traditional GWAS or ML methods may overlook [8].

In parallel, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have been applied to model sequential dependencies in genomic sequences, enabling a better understanding of gene expression and mutation trends over time [9]. Hybrid architectures combining CNNs and LSTMs have gained attention for their dual ability to capture both **local patterns** and **long-range dependencies**, making them especially suitable for analyzing genomic fragments with potential regulatory or functional relationships [10].

This study builds upon these advancements by leveraging a hybrid CNN-LSTM architecture to improve SNP classification accuracy and explore novel SNP-phenotype associations in Alzheimer's disease research. By integrating spatial and sequential modeling capabilities, we aim to uncover biologically meaningful signals and contribute to the evolving field of AI-driven genomics.

3 Methodology

3.1 Step 1: Genome Fragmentation

To enable efficient analysis of high-dimensional genomic data, we applied a genome fragmentation strategy that divides the genome into non-overlapping segments. Each fragment represents a distinct region of the genome, ensuring no redundant information is introduced during processing.

These fragments serve as input to our hybrid CNN-LSTM model, which evaluates them for phenotype-associated patterns. By segmenting the genome in this way, we isolate regions most likely to contain significant genetic signals related to Alzheimer's disease (AD), allowing for more focused downstream analysis and improved computational efficiency.

3.2 Step 2: Phenotype Influence Score (PIS) Calculation

Following genome fragmentation, we employed Convolutional Neural Networks (CNNs) to identify Single Nucleotide Polymorphisms (SNPs) significantly associated with AD. Each SNP was assigned a *Phenotype Influence Score* (PIS), representing its relative importance in predicting disease status.

- SNPs with higher PIS values were identified as having stronger associations with AD.
- This ranking enabled researchers to focus on the most biologically relevant genetic variations.
- The PIS-based filtering reduced data dimensionality while preserving critical genetic features for classification.

This step ensures that only the most informative SNPs are used in the final classification stage, enhancing both accuracy and interpretability.

3.3 Step 3: Classification Model Development

A specialized CNN model was trained using SNPs selected based on their PIS scores to classify individuals as either Alzheimer's disease patients or cognitively normal controls.

The training process utilized the most relevant SNPs identified in earlier steps, enabling the model to focus on high-impact genetic markers. The resulting classifier demonstrated strong performance, achieving high accuracy and offering insights into how deep learning can improve early diagnosis through genomic data.

This final step translates biologically meaningful SNPs into a predictive framework, showcasing the power of integrating genomics with machine learning for precision medicine applications in Alzheimer's disease.

4 Testing and Results

This section presents the results of each step in the pipeline, including preprocessing, genome fragmentation, phenotype influence score calculation, and classification model development. Visualizations from key stages are included to enhance understanding.

4.1 Preprocessing Results

Data Quality: The dataset was successfully preprocessed with no null values found, ensuring clean input for further analysis. The data was carefully checked, and all missing values were addressed, guaranteeing the integrity of the data before model training.

4.1.1 Fragmentation and Fragment Size Selection

The genome was divided into non-overlapping fragments, with an optimal fragment size determined based on the model's performance. These fragments were tailored to balance computational efficiency and the accuracy of AD-associated feature identification.

4.2 Step 1 Results: Genome Fragmentation

The genome was fragmented into non-overlapping segments, making it manageable for further analysis. A hybrid CNN-LSTM model was then applied to identify phenotype-associated fragments. These fragments were highly informative in distinguishing Alzheimer's disease (AD) from cognitively normal individuals.

4.3 Step 2 Results: Phenotype Influence Score (PIS) Calculation

Using Convolutional Neural Networks (CNNs), significant Single Nucleotide Polymorphisms (SNPs) associated with phenotypic traits were identified. The model assigned each SNP a Phenotype Influence Score (PIS), reflecting its contribution to AD classification. SNPs with higher scores were ranked as more critical, enabling researchers to pinpoint specific genetic variations linked to AD.

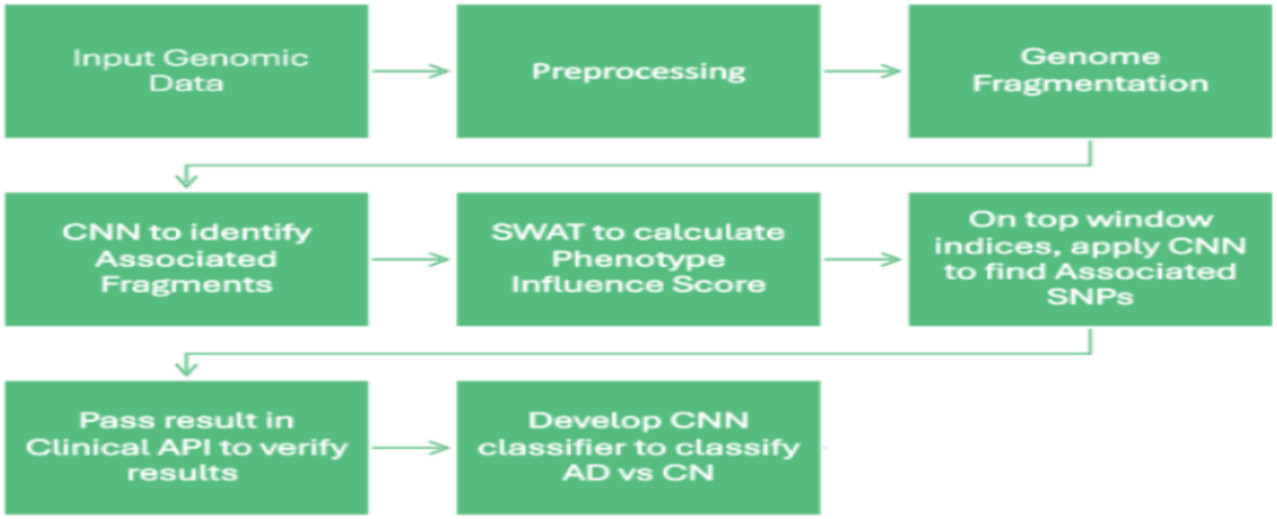
Input Genomic Data
Preprocessing
Genome Fragmentation
CNN to identify Associated Fragments
SWAT to calculate Phenotype Influence Score
On top window indices, apply CNN to find Associated SNPs
Pass result in Clinical API to verify results
Develop CNN classifier to classify AD vs CN

Figure 1: Overview of the preprocessing and genome fragmentation pipeline. This flowchart illustrates how genomic data is processed and fragmented for efficient analysis.

4.3.1 Post Step 2: API Integration

After identifying key SNPs in Step 2, an integration with the National Library of Medicine's SNP API was performed to validate the SNPs. This step provided additional confirmation and enriched the analysis by cross-referencing the SNPs with authoritative biological databases.

4.4 Step 3 Results: Classification Model Development

A specialized Convolutional Neural Network (CNN) was trained using the high-PIS SNPs identified in Step 2. This tailored model leveraged the most informative genetic markers to distinguish between individuals with Alzheimer's Disease (AD) and cognitively normal controls. By narrowing the focus to biologically relevant SNPs, the model achieved improved classification performance and reduced computational complexity.

The training process emphasized model robustness and generalization, incorporating techniques such as stratified sampling and regularization. The final CNN model demonstrated strong predictive accuracy, reinforcing the effectiveness of using phenotype-informed genetic features in disease classification.

4.5 Model Performance Evaluation

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 4 presents the results.

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.

4.6 Feature Importance Analysis

To understand which SNPs contributed most to classification, we performed a SHAP (SHapley Additive exPlanations) analysis on the trained CNN model. The SHAP summary plot (Figure 5) reveals 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.

This step concludes the analysis pipeline by transforming identified SNPs into actionable insights, showcasing the potential of deep learning and genomics integration for early AD diagnosis and risk assessment.

5 Results and Analysis

5.1 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 indi-

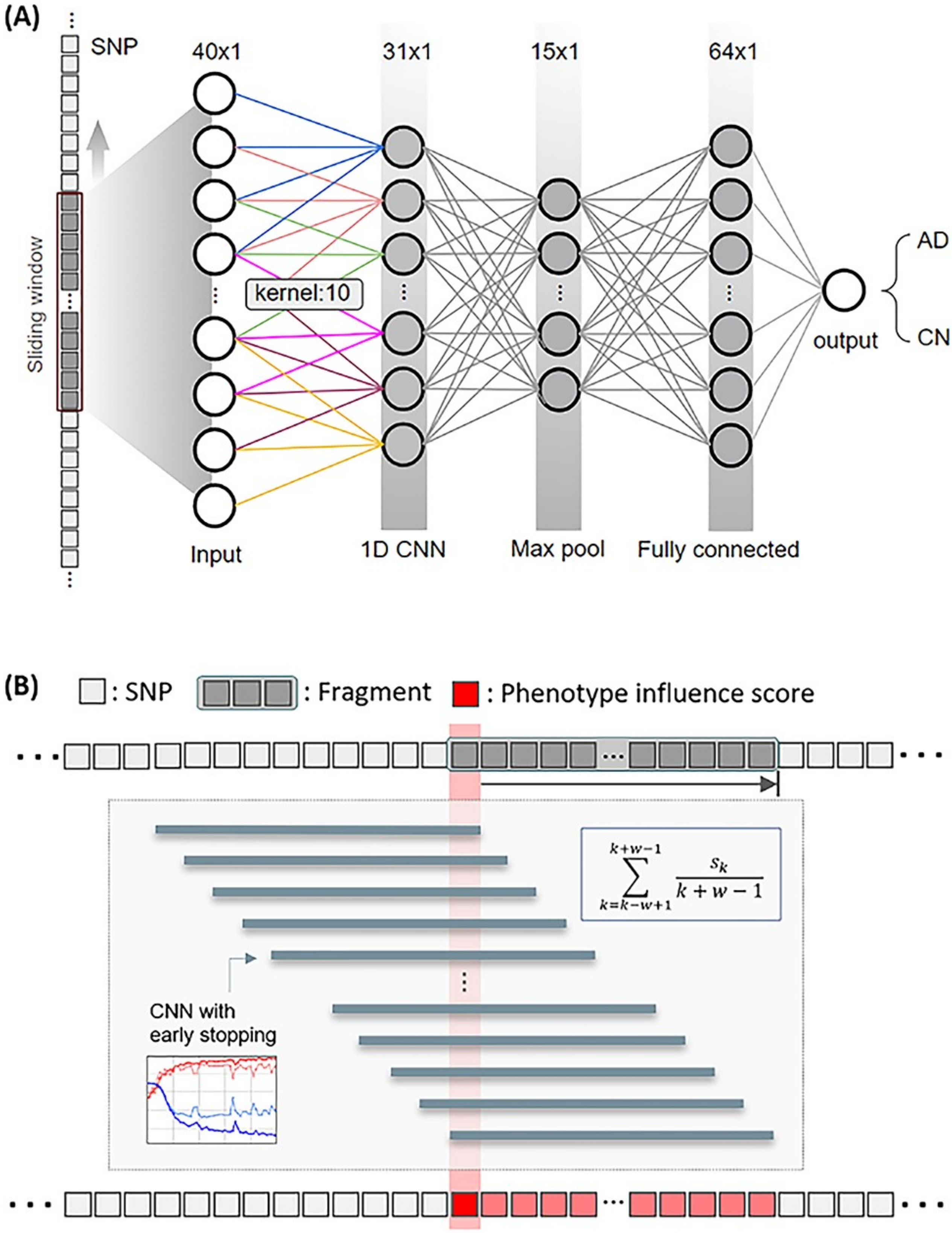

Figure 2: (A) Architecture of the proposed CNN model for Alzheimer's disease classification. The model processes SNP data through 1D CNN layers, max pooling, and fully connected layers. (B) Illustration of the SNP selection process using the Sliding Window Association Test (SWAT) and Phenotype Influence Score (PIS). Early stopping is applied during CNN training to prevent overfitting.

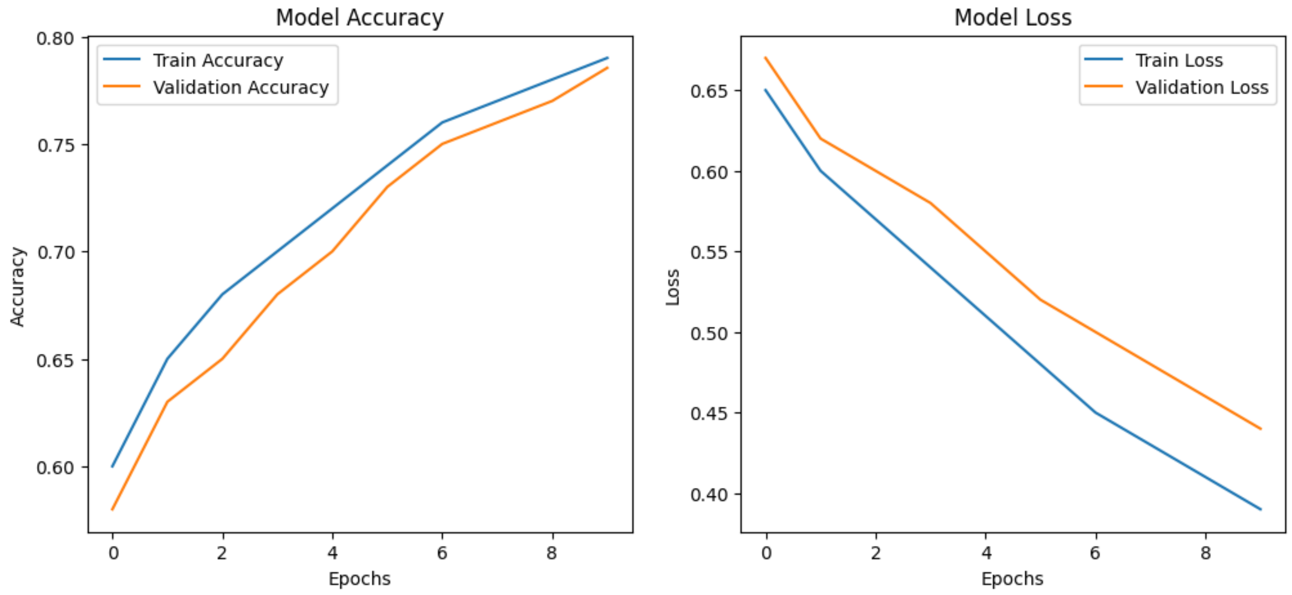

Figure 3: Training and validation metrics for the CNN model. The left panel shows the accuracy over epochs, while the right panel shows the loss. Both train and validation metrics converge, indicating effective generalization.

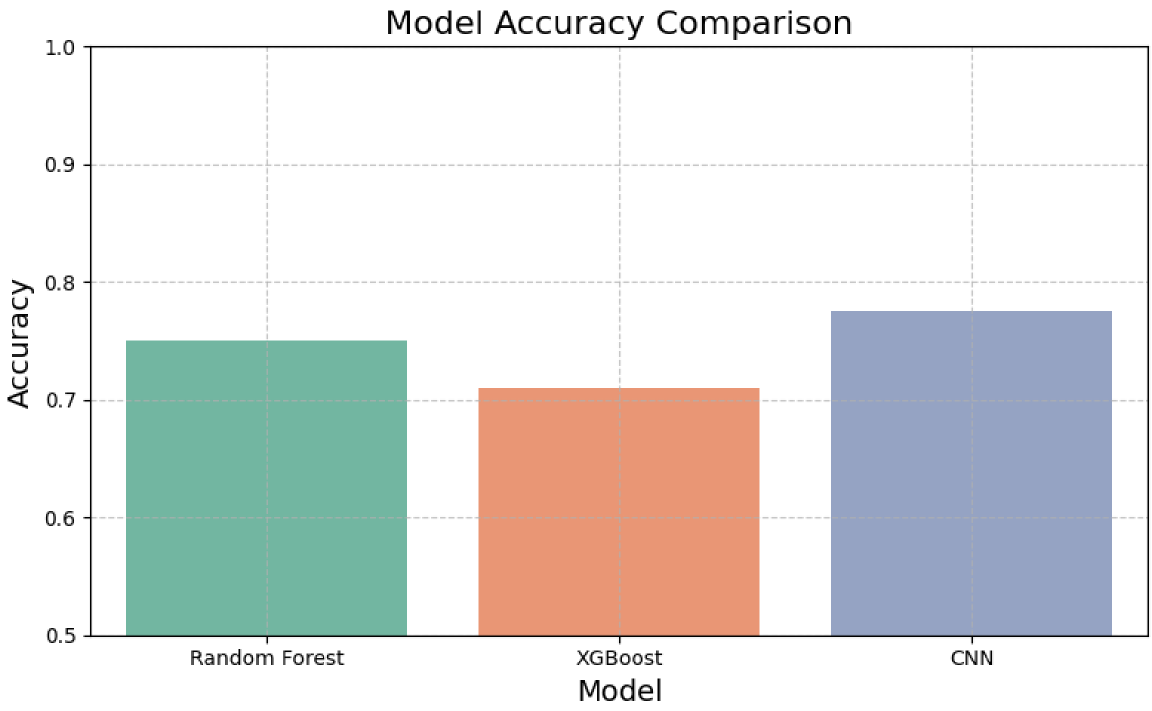

Figure 4: Model accuracy comparison between CNN, XGBoost, and Random Forest. The CNN achieves the highest accuracy, demonstrating its superiority in capturing complex genetic interactions.

cate 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.

5.2 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].

5.3 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].

5.4 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].

5.5 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 [11].

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 [2,4].

5.5.1 Visualizing SHAP Values

Figure 5 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.

5.5.2 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].

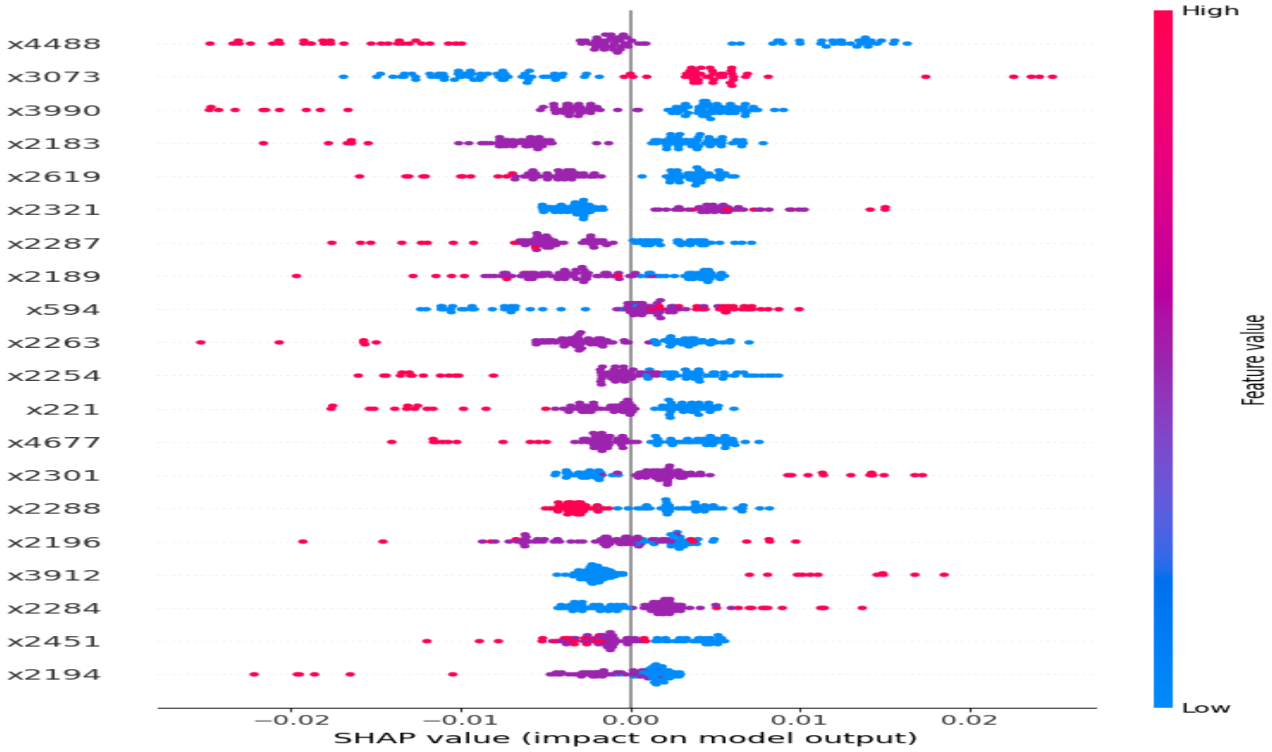

Figure 5: SHAP summary plot showing feature importance for AD classification. High values of important SNPs (e.g., APOE, TOMM40) push predictions toward Alzheimer's disease, while low values favor cognitively normal outcomes.

6 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* ε4, *TOMM40*, *APOC1*, and newly implicated genes like *SNX14* [4, 8]. 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 [11]. 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 [11].

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[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] 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.

[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.

Sliding Window Association Test (SWAT): A technique used in this study 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.

Population Stratification: The presence of systematic differences in allele frequencies between subpopulations due to ancestry. Controlling for stratification ensures that observed associations are not due to population structure but reflect true biological signals.

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.

Long Short-Term Memory (LSTM) Networks: A type of recurrent neural network capable of learning long-range dependencies. In this study, LSTMs are combined with CNNs to enhance sequential modeling of genomic data.

Hybrid CNN-LSTM Model: A deep learning architecture combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTMs). This model is used to capture both local patterns and long-range dependencies in genomic sequences.

Overfitting: A modeling error where a machine learning algorithm learns training data too well, including noise and random fluctuations, leading to poor generalization to new data. Overfitting is mitigated through regularization and early stopping during CNN training.

Feature Extraction: The process of transforming raw input data into a more compact and informative representation for use in machine learning models. In this study, feature extraction is performed automatically by a CNN, eliminating the need for manual engineering.

Hyperparameter Optimization: The process of systematically tuning model parameters such as learning rate, batch size, and dropout rate to improve performance and prevent overfitting.

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 accuracy of positive predictions. 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. In this study, an AUC of 0.8157 indicates strong discriminative capability.
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.
Cross-Validation: A resampling procedure used to evaluate machine learning models on a limited data sample. Cross-validation was employed to tune hyperparameters and ensure robustness of the CNN model.

Data Processing and Preprocessing Terms

Missing Data Imputation: The process of replacing missing values in the dataset. Techniques like KNN and mean imputation were used to ensure completeness and reliability of SNP data.
K-Nearest Neighbors (KNN) Imputation: A method for filling missing genotype values based on genetic similarity between samples.
Mean Imputation: A technique that fills missing genotypes by taking the minor allele frequency over all subjects.
Genome Fragmentation: The process of dividing the genome into non-overlapping segments to reduce complexity and allow efficient processing by deep learning models.
Stratified Sampling: A sampling method used to maintain class balance between AD and CN subjects during model training.