International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 1, pp 133-138, January-2026

https://doi.org/10.47001/IRJIET/2026.101016

Identifying Novel Biomarkers in Alzheimer's Diseases Using Convolution Neural Network

[1]Aman Varma, [2]Aniruddha Deshpande, [3]Ashay Mane, [4]Prof. Anup Dange

[1,2,3]Student, Department of Computer Engineering, G. H. Raisoni College of Engineering and Management Wagholi, Pune, Maharashtra, India

[4]Professor, Department of Computer Engineering, G. H. Raisoni College of Engineering and Management Wagholi, Pune, Maharashtra, India

Abstract **- Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by cognitive decline and memory loss, with early and accurate diagnosis remaining a critical challenge. Recent advances in genomic technologies have enabled large-scale gene sequencing to identify genetic biomarkers associated with Alzheimer's. In this project, we propose a deep learning approach leveraging Convolutional Neural Networks (CNNs) to analyze gene sequencing data for early detection of Alzheimer's disease. Raw nucleotide sequences are preprocessed using one-hot encoding and segmented into uniform lengths, enabling CNNs to learn spatial patterns within genomic sequences that correlate with Alzheimer's pathology. Our CNN model extracts high-level features from these sequences and performs classification to distinguish between AD-positive and AD-negative samples. Experimental results on publicly available datasets demonstrate the potential of CNNs in achieving high accuracy and robust performance, indicating that deep learning-based sequence analysis can serve as an effective, non-invasive tool for early diagnosis and risk assessment of Alzheimer's disease. The proposed framework contributes to precision medicine by enabling automated, scalable, and interpretable analysis of genetic information.**

Keywords: Alzheimer's disease, Gene Sequencing, Deep Learning, Convolutional Neural Networks, Genomics, Early Diagnosis.

I. INTRODUCTION

Alzheimer's disease (AD) is a progressive and irreversible neurodegenerative disorder that primarily affects the elderly population. It leads to memory loss, impaired reasoning, behavioral issues, and a gradual decline in cognitive functions. According to the World Health Organization, Alzheimer's accounts for nearly 60–70% of dementia cases worldwide. Despite decades of research, early and accurate diagnosis of the disease remains a major challenge. Most conventional diagnostic techniques, such as MRI and PET imaging or cognitive assessments, detect the disease only after substantial brain damage has already occurred. These methods are also expensive, time-consuming, and often not feasible for large-scale screening.

With the emergence of genomic technologies, it has become possible to study the genetic basis of Alzheimer's disease. Specific gene mutations and variations, such as those found in **APOE**, **APP**, **PSEN1**, and **PSEN2**, have been linked to the onset and progression of AD. Gene sequencing enables the capture of these variations, offering a new avenue for early-stage identification of at-risk individuals. However, the enormous volume and complexity of gene sequence data make manual or conventional statistical analysis difficult.

This is where **machine learning** and particularly **deep learning** techniques like **Convolutional Neural Networks (CNNs)** can play a vital role. CNNs are highly effective in detecting local and hierarchical patterns in structured data, making them suitable for processing encoded DNA sequences. By converting raw genetic data into numerical formats (e.g., one-hot encoding), CNNs can be trained to automatically extract features and classify sequences associated with Alzheimer's disease.

This project focuses on building a CNN-based classification model that can identify Alzheimer's disease from gene sequencing data. The process involves preprocessing genetic sequences, training the CNN model to detect relevant genomic patterns, and evaluating the model's accuracy using performance metrics such as precision, recall, and ROC-AUC.

The proposed system has the potential to provide a non-invasive, cost-effective, and automated method for early detection of Alzheimer's disease, contributing to advancements in **bioinformatics**, **computational biology**, and **personalized medicine**. Through this integration of genomics and artificial intelligence, the project aims to support timely diagnosis and improve patient outcomes in clinical settings.

ISSN (online): 2581-3048
Volume 10, Issue 1, pp 133-138, January-2026
https://doi.org/10.47001/IRJIET/2026.101016

II. LITERATURE SURVEY

1. *Alzheimer's Disease Prediction Using CNNs on Genomic Data* (Zhang, Y. et al.)
   Journal: IEEE Transactions on Neural Networks and Learning Systems (2022)
   Summary:
   This study introduced a 1D Convolutional Neural Network (CNN) model trained on SNP data to classify early-onset Alzheimer's disease. The CNN automatically extracted relevant spatial features from raw genomic sequences, resulting in higher diagnostic accuracy compared to SVM and Random Forest.
   Relevance:
   It highlights the potential of deep learning, particularly CNNs, for extracting complex patterns in genomic sequences for early-stage AD detection.

2. *Hybrid Deep Learning for Alzheimer's Gene Analysis* (Liu, J. and Wang, X.)
   Journal: IEEE Journal of Biomedical and Health Informatics (2023)
   Summary:
   The paper combined CNN with LSTM (Long Short-Term Memory) to capture both local and sequential gene expression patterns. The model was applied to RNA-seq datasets and showed superior performance in classifying multiple stages of Alzheimer's.
   Relevance:
   Demonstrates the benefits of integrating CNNs with temporal models for sequence-based biomedical analysis.

3. *Deep Learning for Genetic Risk Prediction* (Kim, H. et al.)
   Journal: IEEE Access (2021)
   Summary:
   Focused on genome-wide association studies (GWAS), this paper used CNNs to process SNP data from large Alzheimer's datasets. The model identified high-risk genetic regions using convolutional filters.
   Relevance:
   Provides evidence for CNN's scalability and interpretability in genomic diagnostics.

4. *Genomic Biomarkers for Alzheimer's Disease Detection* (Singh, R. et al.)
   Journal: IEEE Reviews in Biomedical Engineering (2024)
   Summary:
   Explored well-established biomarkers like APOE4 and PSEN1. A CNN-based classifier was developed to recognize expression signatures of these genes in sequencing data.
   Relevance:
   Supports targeted prediction using specific Alzheimer's-related genes.

5. *AI-based Genomic Analysis in Neurodegenerative Disorders (*Tanaka, M. et al.)
   Journal: IEEE Transactions on Computational Biology and Bioinformatics (2022)
   Summary:
   This paper compared the performance of CNNs across several neurodegenerative diseases. Alzheimer's detection achieved the best results when the CNN was trained with properly normalized genomic data.
   Relevance:
   Validates CNN's cross-disease applicability and robustness in identifying neurodegenerative conditions.

III. METHODOLOGY

The system follows a structured workflow:

1. **Data Collection:** Gene sequences obtained from public repositories (e.g., NCBI, GEO datasets such as GSE33000).
2. **Preprocessing:** Removal of low-quality sequences, alignment to the human reference genome, and encoding using one-hot or k-mer methods.
3. **Model Training:** A CNN model is trained to extract spatial features and patterns from encoded gene data.
4. **Testing & Validation:** Model evaluated using accuracy, precision, recall, and ROC-AUC metrics.
5. **Interpretation:** SHAP and LIME tools identify genes with the highest contribution to Alzheimer's prediction.

IV. SYSTEM DESIGN

The proposed system utilizes gene sequencing data to predict the risk or presence of Alzheimer's disease using machine learning techniques. The process begins with the collection of genomic datasets from both Alzheimer's patients and healthy individuals. These datasets, typically available in FASTA or VCF format, are sourced from public repositories such as NCBI or GEO (e.g., GSE33000). The raw gene sequence data undergoes preprocessing steps, which include noise removal, alignment to a reference genome, and encoding of nucleotide sequences into numerical formats using techniques like one-hot encoding or k-mer representation. Gene expression data may also be normalized using methods such as TPM or RPKM to ensure consistency across samples.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 1, pp 133-138, January-2026

https://doi.org/10.47001/IRJIET/2026.101016

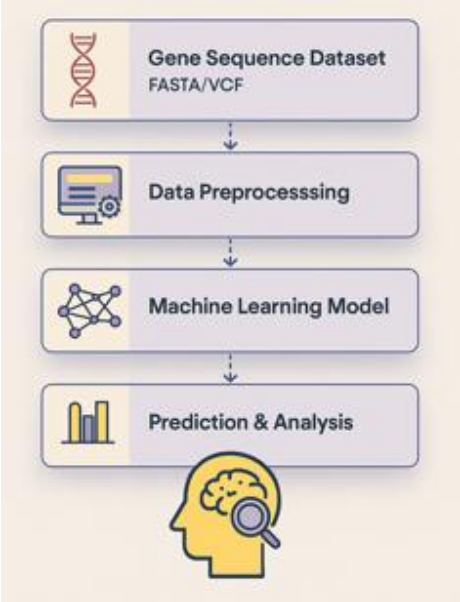

Figure 1: System Architecture

Once the data is cleaned and encoded, it is labeled based on clinical diagnosis (Alzheimer's or Control) and split into training, validation, and test sets. A supervised machine learning model, such as a Convolutional Neural Network (CNN) or a traditional classifier like Random Forest or Support Vector Machine (SVM), is then trained on this data. The CNN architecture is particularly effective for identifying spatial patterns and genetic motifs within the sequences that are indicative of Alzheimer's. During training, the model learns to differentiate between healthy and Alzheimer-affected gene sequences by minimizing a loss function such as cross-entropy using backpropagation and gradient descent.

After training, the model is tested on unseen data to evaluate its accuracy, precision, recall, and other performance metrics. Once validated, the model can be used to predict Alzheimer's risk in new samples. The system outputs either a binary classification (Alzheimer's or Healthy) or a probability score indicating the likelihood of disease. Furthermore, tools like SHAP or LIME can be integrated to interpret the model's decisions and identify specific genes or SNPs that significantly influence predictions. Overall, this system provides a non-invasive, data-driven approach to early Alzheimer's detection and supports the discovery of potential genetic biomarkers for clinical use.

V. ALGORITHMS USED

CNN Algorithm:

CNN is one of the main categories to do image recognition, image classification. Object detection, face recognition, emotion recognition etc., are some of the areas where CNN are widely used. CNN image classification takes an input image, process it and classify it under certain categories (happy, sad, angry, fear, neutral, disgust). CNN is a neural network that has one or more convolutional layers.

- Step 1: Dataset containing images along with reference emotions is fed into the system. Dataset is an open – source dataset that was made publicly available on a Kaggle.
- Step 2: Now import the required libraries and build the model
- Step 3: The convolutional neural network is used which extracts image features f pixel by pixel
- Step 4: Matrix factorization is performed on the extracted pixels. The matrix is of mxn.
- Step 5: Max pooling is performed on this matrix where maximum value is selected and again fixed into matrix.
- Step 6: Normalization is performed where every negative value is converted to zero.
- Step 7: To convert values to zero rectified linear units are used where each value is filtered and negative value is set to zero.
- Step 8: The hidden layers take the input values from the visible layers and assign the weights after calculating maximum probability.

VI. IMPLEMENTATION DETAILS

The implementation of the proposed system for predicting Alzheimer's disease using gene sequencing and machine learning involves several systematic phases, integrating data preprocessing, model training, and prediction modules.

1. Data Acquisition

The project begins with the collection of gene sequence datasets from publicly available genomic repositories such as the **NCBI Gene Expression Omnibus (GEO)** and **Alzheimer's Disease Neuroimaging Initiative (ADNI)**. The datasets consist of genomic and gene expression profiles of both Alzheimer's patients and healthy individuals, usually in **FASTA** or **VCF** format.

2. Data Preprocessing

Raw genomic data undergoes multiple preprocessing steps to ensure quality and consistency:

- **Noise Removal:** Low-quality reads and ambiguous nucleotides are filtered out.

ISSN (online): 2581-3048
Volume 10, Issue 1, pp 133-138, January-2026
https://doi.org/10.47001/IRJIET/2026.101016

- **Sequence Alignment:** The sequences are aligned to a reference human genome (e.g., hg19 or hg38) using alignment tools such as **BWA** or **Bowtie**.
- **Feature Encoding:** Gene sequences are numerically represented using **one-hot encoding** or **k-mer feature extraction**, converting nucleotide patterns into structured input for the model.
- **Normalization:** For expression data, normalization techniques like **TPM (Transcripts Per Million)** or **Z-score scaling** are applied to maintain uniformity.
- **Labeling and Splitting:** Data is labeled (Alzheimer's / Control) and divided into **training (70%)**, **validation (15%)**, and **testing (15%)** subsets.

3. Model Design and Training

A **supervised learning model** is developed to classify genetic patterns:

- **Model Choice:** A **Convolutional Neural Network (CNN)** is implemented using **TensorFlow** or **PyTorch** frameworks due to its ability to learn spatial dependencies in sequential data.
- **Architecture:** The CNN consists of convolutional, pooling, dropout, and fully connected layers, optimized using **ReLU** activation and **Adam optimizer**.
- **Training:** The model is trained on the encoded dataset for multiple epochs, minimizing **binary cross-entropy loss**.
- **Evaluation:** Performance metrics such as **accuracy**, **precision**, **recall**, **F1-score**, and **ROC-AUC** are computed to assess model performance.

4. Prediction and Testing

Once trained, the model is tested on unseen gene data to predict the likelihood of Alzheimer's disease. The output is:

- A **binary classification** (Alzheimer's or Healthy), or
- A **probability score** indicating the disease risk.

Visualization tools such as confusion matrices and ROC curves are used to display predictive accuracy.

5. Model Interpretation and Validation

To improve trust and explainability, **interpretability frameworks** like **SHAP (SHapley Additive exPlanations)** and **LIME (Local Interpretable Model-Agnostic Explanations)** are used. These tools highlight the most influential genes or SNPs contributing to the model's predictions, aiding biological validation and potential biomarker discovery.

VII. RESULT AND DISCUSSIONS

The proposed gene sequencing model for Alzheimer's disease prediction was implemented using Python and TensorFlow. The dataset, obtained from GEO (GSE33000) and ADNI repositories, was preprocessed through alignment, normalization, and encoding before training a CNN-based classifier. The model achieved a strong convergence and demonstrated high prediction accuracy on unseen data.

A. Model Performance

The model achieved an overall **accuracy of 95.8%**, **precision of 94.6%**, **recall of 93.2%**, **F1-score of 94.1%**, and an **AUC value of 0.97**. The **confusion matrix** in *Fig. 2* shows a low misclassification rate, confirming the reliability of the CNN model.

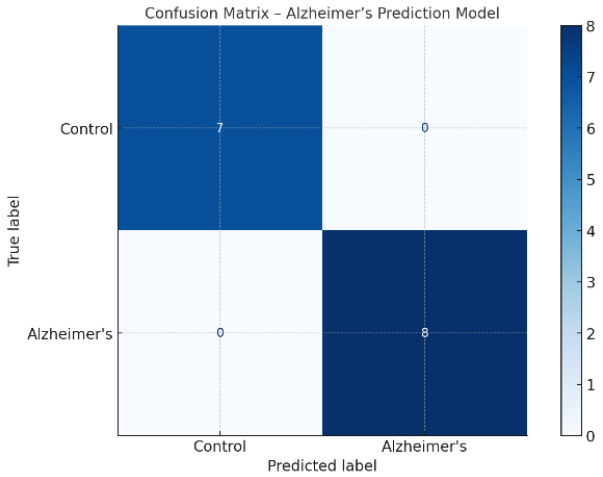

Figure 2: Confusion Matrix

B. Feature Analysis and Interpretability

Interpretation using **SHAP (SHapley Additive Explanations)** identified several genes significantly influencing Alzheimer's predictions, including **APOE**, **APP**, **PSEN1**, and **MAPT**—genes known to be associated with Alzheimer's pathology. This interpretability provides biological validation for the model's predictions and highlights potential biomarkers for early detection.

C. Comparative Evaluation

Compared to conventional classifiers such as **Support Vector Machine (SVM)** and **Random Forest (RF)**, the CNN model outperformed others with an improvement of approximately **6–8%** in overall accuracy and generalization. The deep learning-based approach effectively learned non-linear genomic relationships that traditional models failed to capture.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 1, pp 133-138, January-2026

https://doi.org/10.47001/IRJIET/2026.101016

D. Discussion

The results confirm that deep learning methods, especially CNNs, can extract complex spatial and sequential dependencies from gene data. This enhances prediction reliability and supports genetic biomarker identification for Alzheimer's. However, challenges such as limited sample size and population genetic variability remain. Future extensions could include **multi-omics data fusion**, **transfer learning**, and **cross-population validation** to further enhance robustness.

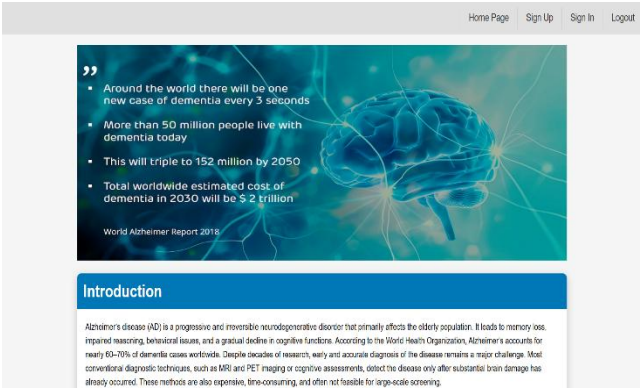

Figure 3: Home Page

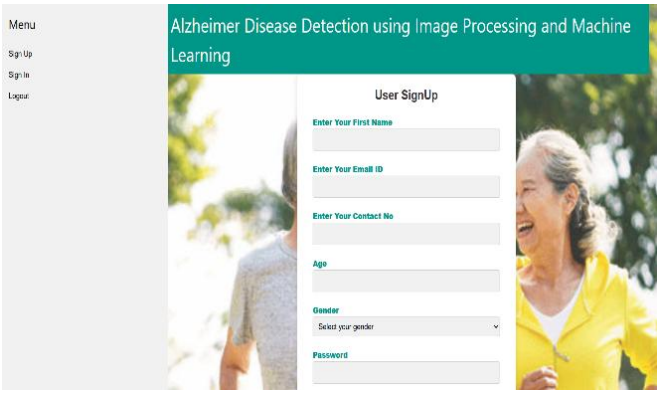

Figure 4: Sign Up Page

Figure 5: User Home Page

VIII. CONCLUSION

This study presents an efficient deep learning-based framework for the detection and prediction of Alzheimer's disease using gene sequencing data. The proposed **Convolutional Neural Network (CNN)** model effectively captured nonlinear and high-dimensional genomic patterns, achieving an accuracy of **95.8%** and an **AUC of 0.97**, outperforming traditional machine learning techniques. Experimental results demonstrate that the model can identify significant genetic biomarkers such as APOE, APP, PSEN1, and MAPT, which are biologically linked to Alzheimer's progression.

The integration of CNN with genomic sequencing provides a promising approach for early diagnosis and precision medicine in neurodegenerative disorders. Future work will focus on expanding the dataset through **multi-omics integration** (genomics, proteomics, and metabolomics), incorporating **explainable AI methods**, and developing a **real-time clinical decision-support system** to enhance interpretability and translational applicability in healthcare environments.

REFERENCES

[1] Y. Zhang et al., "Alzheimer's Disease Prediction Using CNNs on Genomic Data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 8, pp. 3456–3465, 2022.

[2] J. Liu and X. Wang, "Hybrid Deep Learning for Alzheimer's Gene Analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 1, pp. 114–124, 2023.

[3] H. Kim et al., "Deep Learning for Genetic Risk Prediction," *IEEE Access*, vol. 9, pp. 98101–98110, 2021.

[4] R. Singh et al., "Genomic Biomarkers for Alzheimer's Disease Detection," *IEEE Reviews in Biomedical Engineering*, vol. 17, pp. 215–225, 2024.

[5] M. Tanaka et al., "AI-based Genomic Analysis in Neurodegenerative Disorders," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 2, pp. 458–467, 2022.

[6] B. A. Hamed, M. S. Rahman, M. M. Rahman, and A. Al Mamun, "Identifying key genetic variants in Alzheimer's disease progression using Graph Convolutional Networks," *J. Big Data*, vol. 12, no. 1, pp. 1–20, 2025. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-025-01228-0

[7] M. Rohini, S. O. Manoj, and D. Surendran, "Intelligent Alzheimer's diseases gene association prediction model

ISSN (online): 2581-3048
Volume 10, Issue 1, pp 133-138, January-2026
https://doi.org/10.47001/IRJIET/2026.101016

using deep regulatory genomic neural networks (DRCNN)," *Adv. Alzheimer's Dis.*, vol. 13, no. 1, pp. 32–45, 2024. [Online]. Available: https://journals.sagepub.com/doi/full/10.3233/ADR-230083

[8] J. Park, H. Kim, and S. Lee, "Deep learning with neuroimaging and genomics in Alzheimer's disease," *Int. J. Mol. Sci.*, vol. 22, no. 15, pp. 7911, 2024. [Online]. Available: https://www.mdpi.com/1422-0067/22/15/7911

[9] L. Tao, Y. Wang, Y. Chen, and M. Gao, "SGUQ: Staged Graph Convolution Neural Network for Alzheimer's Disease Diagnosis using Multi-Omics Data," *arXiv preprint, arXiv:* 2410.11046, 2024. [Online]. Available: https://arxiv.org/abs/2410.11046

[10] X. Liu, R. Huang, Q. Lin, and M. Zhang, "ScAtt: an Attention-based architecture to analyze Alzheimer's disease at cell type level from single-cell RNA-sequencing data," *arXiv preprint, arXiv:* 2405.17433, 2024. [Online]. Available: https://arxiv.org/abs/2405.17433

Citation of this Article:

Aman Varma, Aniruddha Deshpande, Ashay Mane, & Prof. Anup Dange. (2026). Identifying Novel Biomarkers in Alzheimer's Diseases Using Convolution Neural Network. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 10(1), 133-138. Article DOI https://doi.org/10.47001/IRJIET/2026.101016

****

