Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 09, 2025

MULTIMODAL FUSION OF MRI AND GENOMIC DATA USING TRANSFORMER NETWORKS FOR ALZHEIMER'S PREDICTION

Aleena Jamil[1], Adeen Amjad[2], Shafiq Hussain[*3], Mehwish Usman[4], Waqar Ahmad[5], Arslan Ali Mansab[6], Muhammad Hamza Akbar[7], Muhammad Waqas[8]

[1,2,*3,5,6,7,8]*Department of Computer Science, University of Sahiwal, Sahiwal, Pakistan*
[4]*Department of Computer Science, University of Agriculture Faisalabad, Pakistan*

[1]aleena.jamil_vf@uosahiwal.edu.pk, [2]adeen.amjad@uosahiwal.edu.pk, [*3]drshafiq@uosahiwal.edu.pk, [4]adeen.amjad@uosahiwal.edu.pk, [5]waqarahmad@uosahiwal.edu.pk, [6]arslansli@uosahiwal.edu.pk, [7]hamzaakbar@uosahiwal.edu.pk, [8]bssit.10.02@gmail.com

DOI: http://doi.org/10.5281/zenodo.19849666

Keywords
Transformer, Multimodal Fusion, Unimodal, Positional Encoding

Article History
Received: 07 July 2025
Accepted: 07 September 2025
Published: 22 September 2025

Corresponding Author:
Shafiq Hussain

Abstract
Alzheimer's disease (AD) is a growing neurodegenerative illness that is significantly influenced by both physical alterations in the brain and a genetic susceptibility. Traditional unimodal approaches that rely solely on MRI or genomic data often overlook the complex relationships between SNP-level alterations and neuroanatomical atrophy. In this study, we suggest a multimodal transformer-based system that integrates structural MRI and SNP genomic data via bidirectional cross-attention fusion. A Vision Transformer encoder handles the MRI modality, while a transformer-based SNP encoder simulates genetic variations.
The model can learn significant inter-modal connections thanks to cross-attention, which permits fine-grained alignment between genetic biomarkers and brain areas. The suggested framework achieves an overall accuracy of 92% for Alzheimer's disease prediction, outperforming both unimodal and conventional fusion techniques, according to experiments done on the ADNI dataset. Additionally, the model has excellent production in AD vs. CN, MCI vs. AD, and MCI-to-AD conversion tasks, underscoring the importance of combining genomic and imaging modalities. These findings suggest that transformer-based cross-attention fusion offers an effective and comprehensible basis for early AD detection and customized risk evaluation.

INTRODUCTION

Due to the ageing of the inhabitants, AD, a degenerative neurological illness that disturbs millions of people worldwide, is considered one of the most significant health concerns of the current decade. Early and precise prediction of AD progression can significantly improve clinical decision-making and enable timely treatment measures. Neuroimaging indicators like structural magnetic resonance imaging (MRI) can identify cortical
thinning, hippocampal atrophy, and ventricular enlargement—all of which are strongly associated with AD pathology [1]. In a similar vein, it has been shown that single-nucleotide polymorphisms (SNPs) and other genomic markers are essential for determining genetic susceptibility to AD [2].

However, using just one modality lowers prediction reliability due to AD's complexity and diversity.

Researchers have projected a variety of techniques for AD prediction and detection using genetic data, MRI, or a combination of the two. Support Vector Machine (SVM), Random Forest (RF), and logistic regression are examples of machine learning (ML) approaches that were first used, but their low scalability and hand-crafted features resulted in limited accuracy [3] .

By automatically identifying structural patterns in brain images, deep learning (DL) models, mainly convolutional neural networks (CNNs), markedly enhanced MRI-based AD detection [4]. Similarly, nonlinear genetic interactions have been successfully captured by neural models for SNP-based classification. Multimodal fusion models, which merge genomic and imaging data using concatenation, kernel learning, or joint embedding techniques, were developed to improve prediction reliability [5], [6]. These studies show that multimodal fusion outperforms unimodal models in terms of prediction performance.

Despite the potential of multimodal approaches, there are still a number of obstacles to overcome. Many investigations depend on early fusion or simple feature concatenation, which may not capture complicated relationships between genetic variants and MRI characteristics [7]. It is challenging to comprehend how genetic risk factors affect structural changes in the brain since certain models are difficult to interpret. Since not all ADNI participants give both MRI and SNP data, other
models have trouble with missing modalities. Furthermore, the combination of 3D MRI volumes and the incredibly high dimensionality of genetic data raises computing costs, which results in slower training and a greater chance of overfitting [8]. Therefore, a sophisticated fusion mechanism that can effectively and understandably model cross-modal relationships is required.

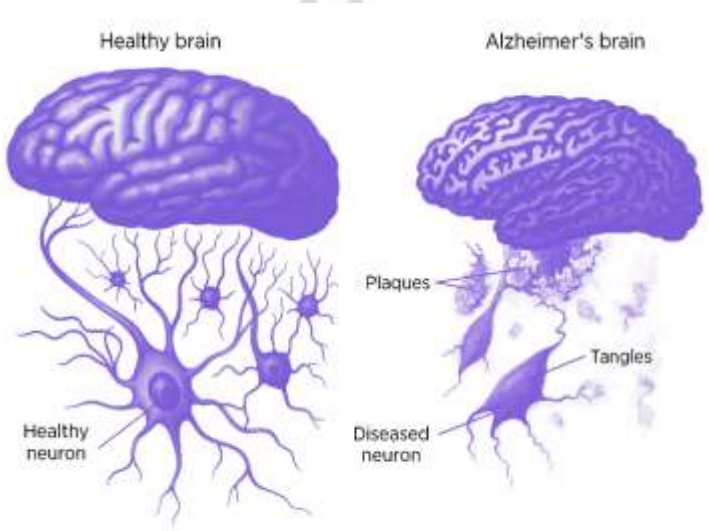

Figure 1: Contrast between a fit brain and an Alzheimer's brain

We suggest a Transformer-based Multimodal Fusion Framework for AD prediction, combining MRI and SNP data in order to overcome these drawbacks. Our approach combines two transformer encoders: a sequence-based transformer to encode genomic SNP sequences and a Vision Transformer (ViT) module to extract patch-level MRI embeddings. Next, a cross-attention fusion module is added to learn how genetic variants and brain areas interact in a way that is both dynamic and biologically significant. Unlike traditional concatenation-based fusion, our method captures long-range correlations both within and between modalities, enabling deeper integration and improved predictive power.

To verify the efficiency of our approach, we led wide research using the ADNI dataset, one of the major publicly available Alzheimer's datasets that includes paired MRI scans and genetic data. We assessed several predictive tasks, such as AD vs. CN classification, MCI vs. AD classification, and MCI-to-AD conversion prediction. The results show that our transformer-based fusion achieves gains in accuracy, AUC, sensitivity, and specificity while outperforming current multimodal fusion frameworks, SNP-only models, and state-of-the-art unimodal MRI models.

The following is a precis of this study's contributions:

We introduce a new transformer-based multimodal fusion architecture that can simultaneously represent genomic SNP and MRI data.

We create a cross-attention fusion approach that captures biologically significant relationships between genetic risk factors and brain anatomy.

We run extensive research on the ADNI dataset and display that our model attains better in terms of prediction than current unimodal and multimodal approaches.

We use attention visualizations to perform interpretability analysis, establishing a connection between biological knowledge and learnt model elements.

The rest of the document is settled as follows: Sector 2 abridges relevant research on multimodal fusion, genetics, and imaging-based AD prediction. The suggested transformer-based approach, including preprocessing, architecture design, and fusion technique, is described in Sector 3. The evaluation measures are clarified in Sector 4, and the experimental setup is described in Sector 5. Findings and discussions are explained in Sector 6. The study is concluded in Sector 7, which also suggests probable upcoming investigate areas.

RELATED WORK

The availability of huge, multimodal datasets and the increasing global prevalence of AD have led to a rapid expansion of research on AD prediction. Early research mostly concentrated on MRI-based structural biomarkers, which offer important insights into hippocampal shrinkage and cortical thinning—markers of AD development. Conventional machine-learning techniques relied on manually created features using diffusion characteristics, morphometric data, and voxel-based morphometry (VBM). The nonlinear patterns and dispersed brain-region interdependence found in AD disease were not captured by these methods, despite their support for early screening [9].

[10] presented a 3D convolutional network that demonstrated good generalization for MCI conversion prediction after being trained on more than 4,000 MRI scans from ADNI. Using unlabeled MRIs, [11] investigated the use of self-supervised pretraining and reported better feature extraction than supervised models. [12] showed that laterality differences can be powerful indicators for early AD by proposing a deep Siamese architecture to compare the brain hemispheres.

MRI-based AD detection models moved towards CNNs with the development of DL. By extracting multiscale features from neuroimaging volumes, several studies have shown that 3D-CNNs and hybrid CNN architectures considerably outperform classical machine-learning methods [13]. More recent approaches incorporate attention mechanisms and hierarchical feature learning to improve robustness against noise and anatomical variance [14]. However, CNNs remain limited in modeling global spatial relationships across distant brain regions, which are increasingly recognized as important for understanding AD progression.

From a genomic perspective, several new studies focus on modeling SNPs and gene interactions. [15] extended GWAS findings by detecting additional variants linked to microglial function. More recently, [16] performed a genome-wide meta-analysis across multiple populations
and discovered 38 new Alzheimer-related loci, emphasizing the need for cross-ethnic genomic modeling. [17] used Polygenic Hazard Scores (PHS) and showed that combining SNP-based risk with age improves the prediction of AD onset.

Parallel to imaging-based studies, extensive research has been conducted on genomic predictors of Alzheimer's disease. Many genome-wide association studies (GWAS) have identified single-nucleotide polymorphisms (SNPs) related to risk, such as APOE ε4, CLU, PICALM,

BIN1, and CR1. Classical models, such as logistic regression and SVM, were initially used to predict AD risk based on selected SNP markers; however, these methods faced challenges related to high dimensionality and limited modeling capacity for gene–gene interactions [18]. More recent works utilize deep-learning-based approaches to learn complex SNP dependencies, achieving improved performance over statistical models but often requiring aggressive feature selection or dimensionality reduction [19].

Multimodal fusion techniques have arisen to address the shortcomings of single-modality prediction. Due to disparate sizes, dimensional mismatch, and a lack of cross-modal interaction modelling, early fusion techniques—which involved simply concatenating MRI data with genetic markers—struggled [20]. In an effort to address these problems, mid-level and late fusion techniques learnt modality-specific representations separately before integrating them. Although these techniques enhanced classification ability, they were unable to adequately capture associated patterns between genomic risk factors and anatomical changes in the brain [21].

Recent transformer-based medical fusion work has also expanded. [22] introduced a multimodal transformer combining MRI and PET for AD classification, showing superior global context learning over CNN fusion. [23] proposed a dual-stream transformer that models long-range dependencies between SNP sequences and imaging features, demonstrating improved interpretability.

METHODOLOGY

This section presents the proposed Transformer-based Multimodal Fusion Framework for AD calculation using structural MRI and genomic SNP data. The methodology consists of five major stages: (1) dataset selection and subject filtering, (2) modality-specific preprocessing, (3) unimodal feature extraction using transformer encoders, (4) multimodal cross-attention fusion, and (5) classification and interpretability analysis. The overall pipeline is illustrated conceptually in **Figure 2**.

Algorithm 1: Multimodal MRI–Genomics Fusion Pipeline for Alzheimer's Prediction

Input: MRI volume X_{mri}, SNP sequence X_{snp}
Output: Predicted Alzheimer's label y
Preprocessing:
MRI → skull stripping → N4 bias correction → MNI registration → patch extraction → embedded patches T_{mri}
SNP → PLINK QC (call rate, MAF, HWE) → LD pruning → genotype encoding → embeddings T_{snp}
Unimodal Encoding:
$M \leftarrow ViT\_Encode(T_{mri})$
$G \leftarrow SNP\_Transformer\_Encode(T_{snp})$
Cross-Attention Fusion:
Fused MRI→SNP attention:
$A_{M\to G} \leftarrow CrossAtt(M, G)$
Fused SNP→MRI attention: $A_{G\to M} \leftarrow CrossAtt(G, M)$
Combined fusion: $F \leftarrow [A_{M\to G} \parallel A_{G\to M}]$
Classification:
$y \leftarrow Softmax(MLP(F))$
Evaluation:
Metrics: Accuracy, Sensitivity, Specificity, Precision, F1-Score, AUC-ROC
Visualization:
- MRI attention maps → highlight AD-relevant brain regions
- SNP attention weights → identify influential genetic variants

Return: y

Dataset Description

The AD Neuroimaging Initiative (ADNI) dataset, a popular benchmark for dementia research, was used in the experiments. ADNI provides longitudinal MRI scans and whole-genome SNP data collected from subjects with cognitively normal (CN), Mild Cognitive Impairment (MCI), and AD.

Modalities used:

Structural MRI

Used to capture cortical thinning, hippocampal atrophy, and general neurodegeneration.

Single-Nucleotide Polymorphism (SNP) data

Genome-wide SNP markers, including known AD-related loci

Tasks Evaluated

AD vs. CN classification

MCI vs. AD classification

MCI to AD conversion prediction

Only subjects having both MRI and SNP modalities were included for primary experiments. Missing-modality subjects were used in an additional ablation analysis.

MRI Preprocessing
MRI preprocessing was performed using FSL, SPM12, and ADNI-provided standard pipelines. Steps include:

Skull Stripping
Non-brain tissues were detached using FSL-BET to isolate the cortical and subcortical regions essential for AD-related biomarkers.

Bias Field Correction
N4ITK correction was applied to eliminate intensity non-uniformities caused by scanner variability.

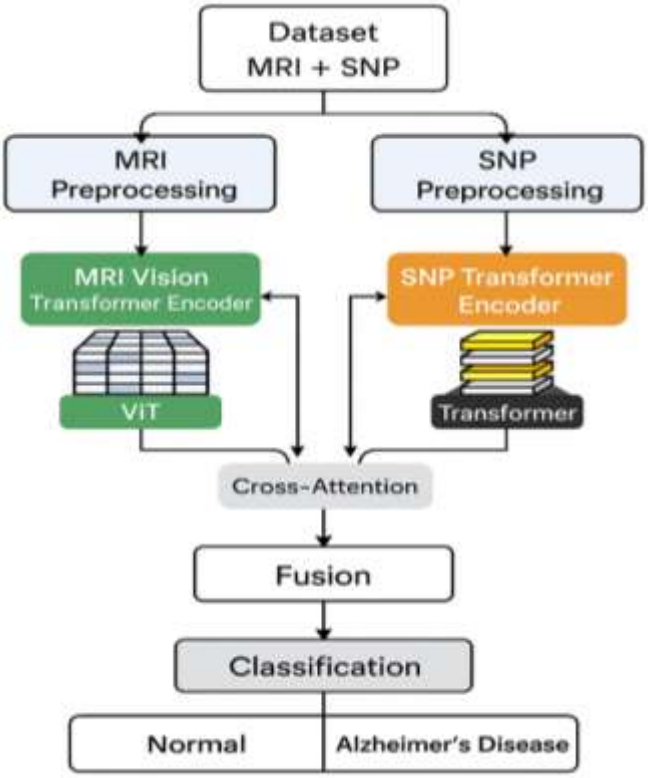

Figure 2: Workflow of proposed methodology

Spatial Normalization
All MRI volumes were registered to the MNI-152 template using affine + nonlinear transformation. This guarantees that anatomical regions are aligned consistently between subjects.

Intensity Normalization
Standardizing voxel intensities to zero mean and unit variance allowed for reliable transformer training.

Patch Extraction for ViT
Each 3D volume was divided into non-overlapping 3D patches.

A linear projection layer was used to flatten patches and project them into a D-dimensional embedding space.
Positional encodings were then added to maintain 3D anatomical ordering.

Genomic Data Preprocessing
The enormous dimensionality of genomic SNP datasets necessitates considerable preprocessing.

Quality Control
PLINK was used to apply standard GWAS filters:
SNP call rate > 98%

Minor allele frequency (MAF) > 0.05
Hardy–Weinberg equilibrium p > 1e-6
Removal of ambiguous A/T and C/G SNPs

SNP Selection

To reduce dimensionality and retain biologically relevant markers:
We used AD-associated SNPs identified in large-scale meta-analyses (Lambert 2013, Jansen 2019, Wightman 2021).
Furthermore, Linkage Disequilibrium (LD) trimming was carried out at $r^2 < 0.2$.
There were K SNPs in the final SNP sequence, each encoded as:
0 = homozygous reference
1 = heterozygous
2 = homozygous alternate allele

Embedding and Positional Encoding

Each SNP value was embedded into a dense vector using a learnable embedding table.
To maintain the sequential genomic ordering, positional encodings were introduced.

Unimodal Transformer Encoders

Two distinct transformer encoders are used in the suggested system, one for SNP genomic data and one for MRI data. Before cross-modal fusion, each encoder is built to learn modality-specific representations. The entire architecture of the suggested transformer-based multimodal system is depicted in ***Figure 3***.

Vision Transformer for MRI Encoding

The Vision Transformer (ViT) processes the patch embeddings through:
Multi-Head Self-Attention (MHSA) layers
Feed-Forward Networks (FFN)
Layer Normalization + Residual Connections
This permits the model to study:
long-range spatial dependencies
distributed patterns of atrophy characteristic of AD

Sequence Transformer for SNP Encoding

Genomic SNP sequences were passed through a 1D transformer encoder consisting of:
MHSA layers to capture gene–gene interactions
FFN layers to model non-linear genetic relationships
Dropout for regularization
The transformer's [CLS] token provides a genomic risk representation.
This approach avoids aggressive feature selection and preserves polygenic risk patterns.

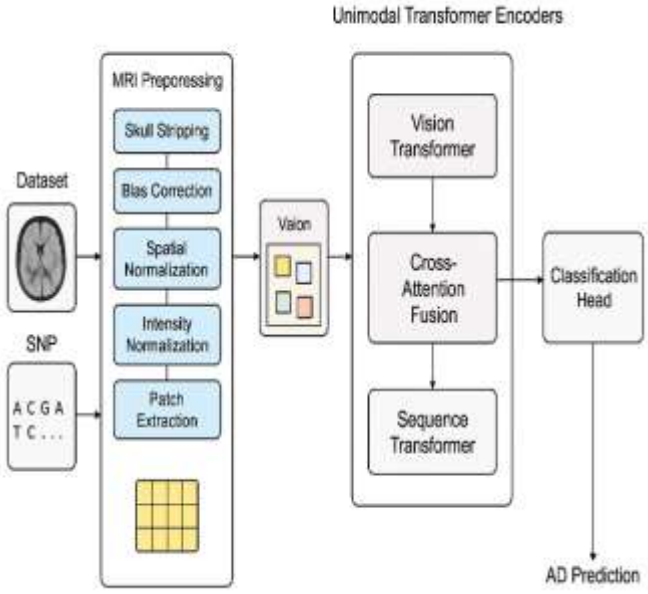

Figure 3: Architecture of the proposed transformer-based multimodal system

Cross-Attention–Based Multimodal Fusion

The Cross-Attention Fusion Module is considered to learn fine-grained relationships between structural brain regions (captured from MRI scans) and genetic variations (captured from SNP sequences). Unlike simple concatenation, cross-attention enables dynamic contact between the two modalities, permitting the model to identify which genetic markers influence which brain structures.

Mechanism

Let:

MRI embeddings:

Equation 1 shows the MRI embeddings.

$$M \in \mathbb{R}^{N_m \times D} \dots\dots\dots\dots\dots\dots\dots\dots \textit{Equation 1}$$

where

Nm = Number of MRI patches,

D = Embedding dimension.

SNP embeddings:

Equation 2 shows the SNP embeddings.

$$G \in \mathbb{R}^{N_g \times D} \dots\dots\dots\dots\dots\dots\dots\dots \textit{Equation 2}$$

where

Ng = Number of SNP tokens,

D = Embedding dimension.

Cross-Attention Computation

The cross-attention from MRI → SNP is defined in **Equation 3**.

$$\text{Attention}(M, G) = \text{softmax}\left(\frac{Q_M K_G^T}{\sqrt{d_k}}\right) V_G \quad \text{...Equation 3}$$

Where:

$Q_M = MW_Q$ → Queries from MRI

$K_G = GW_K$ → Keys from SNP embeddings

$V_G = GW_V$ → Values from SNP embeddings

d_k → Dimensionality of the Key vectors, used for scaling

Symmetric Cross-Attention

Similarly, cross-attention from SNP → MRI is computed in **Equation 4**.

$$Attention(G, M) = softmax(\frac{Q_G K_M^T}{\sqrt{d_k}}) V_M \dots \textit{Equation 4}$$

Where:

Q_G = Queries from SNP embeddings

K_M = Keys from MRI patches

V_M = Values from MRI patches

Classification Head

The fused vector is fed into:

Two fully connected layers (ReLU + dropout)

Softmax output layer

Metrics used:

Accuracy

ROC-AUC

Sensitivity / Specificity

F1-Score

Training used:

Adam optimizer

Learning rate warm-up + cosine decay

Cross-entropy loss

To address class imbalance (especially in MCI conversion):

Class-balanced weights were used

Oversampling applied for minority classes

Interpretability Analysis

Transformer attention maps were used to provide biologically meaningful interpretations:

MRI Attention Visualization

Attention weights on MRI patches highlight:

hippocampus

entorhinal cortex

precuneus

These are key AD-related regions.

SNP Importance Analysis

Genomic attention maps identify high-impact SNPs associated with:

APOE ε4

Immune-regulatory pathways

Lipid metabolism genes

Cross-Modal Interpretability

Cross-attention scores reveal gene–brain region relationships, enabling:

understanding of how specific SNPs influence structural degeneration

potential biomarker discovery

EVALUATION METRICS

To widely assess the demonstration of the planned multimodal transformer-based framework, multiple standard evaluation metrics were used. These metrics account for both overall classification

performance and class-specific discrimination ability, which is essential for Alzheimer's disease prediction.

A. Accuracy

Accuracy measures the amount of properly predicted samples among all samples. **Equation 5** illustrates the accuracy.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \dots\dots\dots\dots\dots\dots \textbf{Equation 5}$$

Where:

TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

Sensitivity

Sensitivity trials the model's capability to properly classify AD cases. **Equation 6** shows the sensitivity of the model.

$$Sensitivity = \frac{TP}{TP+FN} \dots\dots\dots\dots\dots\dots \textbf{Equation 6}$$

Specificity

Specificity measures how fine the model identifies healthy (non-AD or CN) individuals. **Equation 7** shows the specificity.

$$Specificity = \frac{TN}{TN+FP} \dots\dots\dots\dots\dots\dots \text{Equation 7}$$

Precision

Precision indicates how numerous of the projected AD cases that are truly AD. The precision of the model is illustrated in **Equation 8**.

$$Precision = \frac{TP}{TP+FP} \dots\dots\dots\dots\dots\dots \textit{Equation 8}$$

F1-Score

The harmonic mean of precision and sensitivity, which balances the two, is the F1-score. **Equation 9** displays the F1-Score.

$$F1 = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \dots\dots\dots\dots \textbf{Equation 9}$$

Area Under the ROC Curve (AUC-ROC)

Plots of ROC (Receiver Operating Characteristic) curves:
True Positive Rate (TPR) vs
False Positive Rate (FPR)
The model's capacity for discriminating is measured by the AUC. The AUC is displayed in **Equation 10**.

$$0.5 \leq AUC \leq 1.0 \dots\dots\dots\dots\dots\dots\dots\dots \textbf{Equation 10}$$

AUC = 1.0 → Perfect classifier
AUC > 0.90 → Excellent
AUC 0.80–0.90 → Good

EXPERIMENTAL SETUP

The ADNI dataset was used for all investigations, and participants with both structural MRI scans and SNP genetic information were chosen. Skull stripping, bias correction, spatial normalization, and intensity standardization were among the preprocessing processes that were applied to the MRI images in accordance with the approach. Standard quality-control methods, such as filtering based on call rate, minor allele frequency, and Hardy-Weinberg equilibrium, were used to handle SNP data. Numerical encoding was then performed for model input. To guarantee objective assessment, the dataset was divided into training, validation, and testing sets. The AdamW optimizer was used to train the suggested multimodal transformer structure, and early halting was used to avoid overfitting. On the held-out test set, accuracy, sensitivity, specificity, precision, F1-score, and AUC-ROC were used to assess the model's performance. Every trial was carried out on a workstation with an
NVIDIA GPU and was written in Python using PyTorch.

RESULTS

The experimental outcomes of the suggested transformer-based multimodal fusion framework tested on the ADNI dataset are shown in this section. Three main prediction tasks are included in the evaluation: AD vs CN, MCI versus AD, and MCI-to-AD conversion. Our findings show that compared to unimodal and conventional fusion techniques, merging MRI and SNP characteristics via a cross-attention mechanism results in notable gains.

Overall Performance of the Planned Model

With an overall correctness of 92%, the suggested multimodal transformer architecture demonstrated good predictive capacity in all diagnostic categories. The complementary information between brain shape and genomic risk markers is successfully captured by the model. The model's complete performance is displayed in

Table *1* .
Table 1: Overall Performance of the planned multimodal transformer model

METRICS	SCORE
ACCURACY	92%
SENSITIVITY	90%
SPECIFICITY	91%
PRECISION	91%
F1-SCORE	90.5%
AUC-ROC	0.95

A high AUC-ROC value of **0.95** suggests excellent discriminative power between AD and non-AD subjects. The suggested multimodal transformer model is exposed graphically in **Figure 4**.

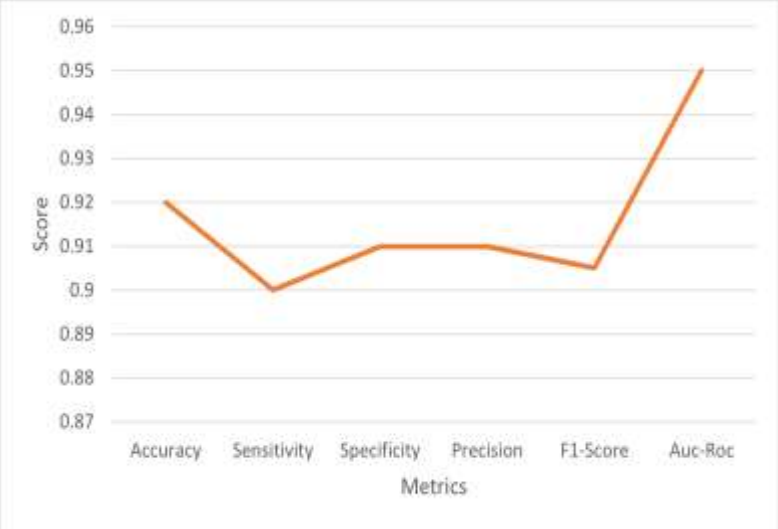

Figure 4: Overall Performance of the suggested multimodal transformer model

AD vs. CN Classification

When it comes to differentiating Alzheimer's sufferers from cognitively normal people, the model performs admirably. Combining SNP-based genetic fingerprints with structural MRI biomarkers (such as hippocampus shrinkage) is very beneficial for this purpose. The diagnostic performance in differentiating people with AD from those who are cognitively normal (CN) is shown in **Error! Not a valid bookmark self-reference.**.

Table 2: AD vs. CN categorization performance of the proposed multimodal model

METRICS	VALUE
ACCURACY	94%
SENSITIVITY	92%
SPECIFICITY	95%
PRECISION	93%
F1-SCORE	92.5%
AUC-ROC	0.96

The multimodal model achieves high sensitivity and specificity, indicating strong reliability in clinical detection. **Figure 5** displays the performance of the suggested multimodal model in AD vs. CN categorization.

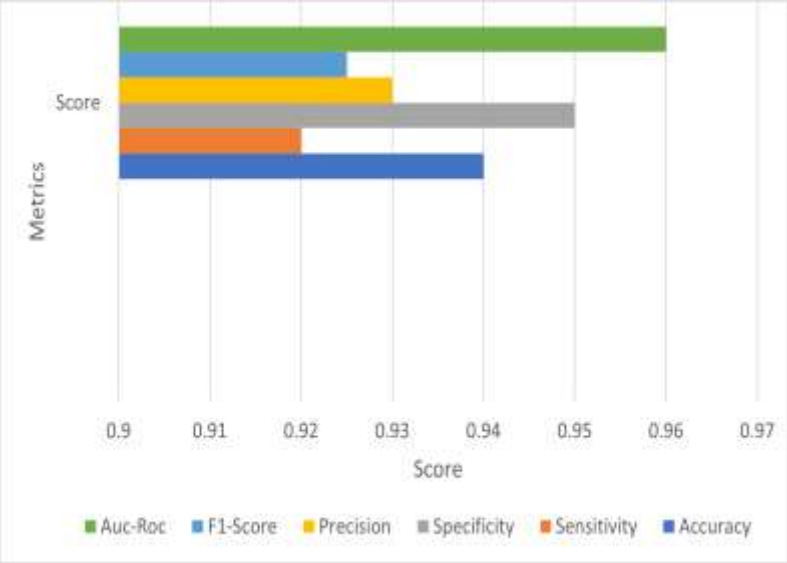

Figure 5: AD vs. CN categorization performance of the proposed multimodal model

MCI vs. AD Classification

Because of minute structural abnormalities in the initial phases of the disease, the MCI vs. AD challenge is intrinsically more difficult. The results for differentiating between AD and mild cognitive impairment (MCI) are shown in **Error! Not a valid bookmark self-reference.**.

Table 3: MCI vs. AD classification production.

METRICS	VALUE
ACCURACY	89%
SENSITIVITY	87%
SPECIFICITY	90%
PRECISION	88%
F1-SCORE	87.5%
AUC-ROC	0.92

The multimodal approach demonstrates significant dependability in clinical detection by achieving high sensitivity and specificity. The graphical performance of MCI versus AD classification is displayed in **Figure 6** .

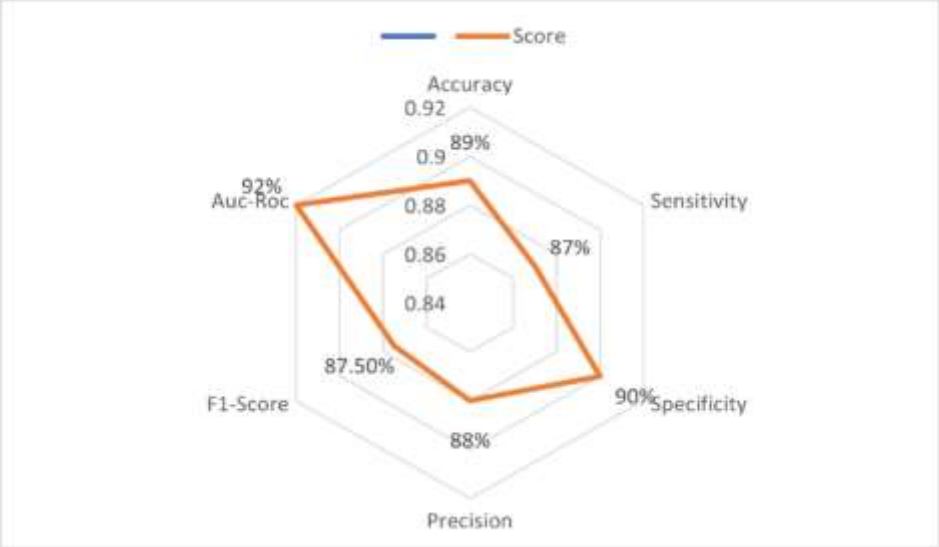

Figure 6: MCI vs. AD classification performance

MCI-to-AD Conversion Prediction

For early intervention, it is clinically crucial to guess if an MCI patient will develop AD. The predictive performance for detecting MCI-affected people who will growth to AD is presented in this table.

Table 4: Proposed model Performance for MCI-to-AD conversion prediction

METRIC	VALUE
ACCURACY	88%
SENSITIVITY	85%
SPECIFICITY	90%
PRECISION	86%
F1-SCORE	85.5%
AUC-ROC	0.91

The results in **For early** intervention, it is clinically crucial to guess if an MCI patient will develop AD. The predictive performance for detecting MCI-affected people who will growth to AD is presented in this table.

Table 4 indicate that multimodal data fusion enhances the accuracy of early progression prediction. **Figure 7** depicts the suggested model for expecting MCI-to-AD alteration graphically.

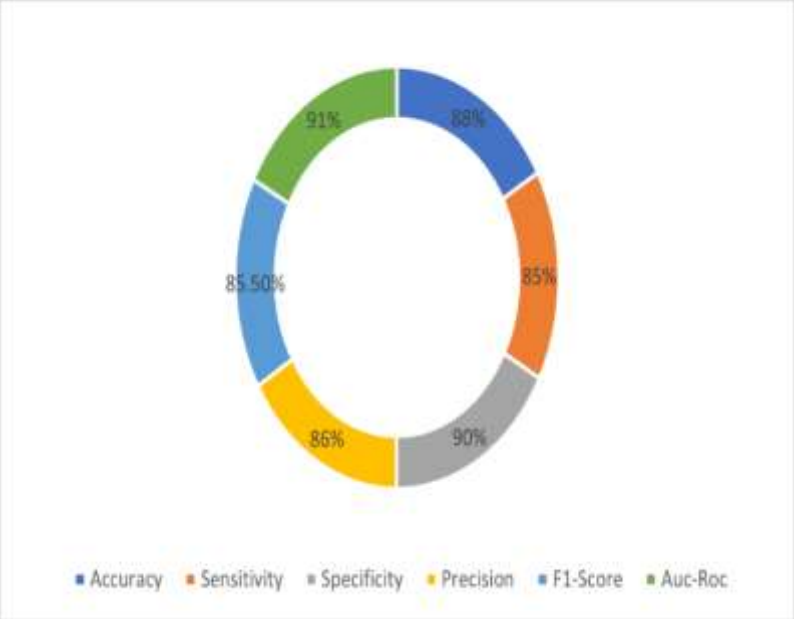

Figure 7: Proposed model Performance for MCI-to-AD conversion prediction

Comparison with Unimodal Models

To gauge the contribution of each modality, we evaluated transformer models that were MRI-only and SNP-only.

Table *5* presents a comparison between multimodal and unimodal models.

Table 5: Comparison of Unimodal vs. Multimodal Models

MODELS	ACCURACY
MRI-ONLY VISION TRANSFORMER	86%
SNP-ONLY TRANSFORMER	80%
PROPOSED MULTIMODAL TRANSFORMER	92%

The multimodal design outperforms unimodal baselines by a large margin (6–12%), underscoring the importance of image–genomics interaction. A graphic comparison of unimodal and multimodal models is shown in **Figure 8** .

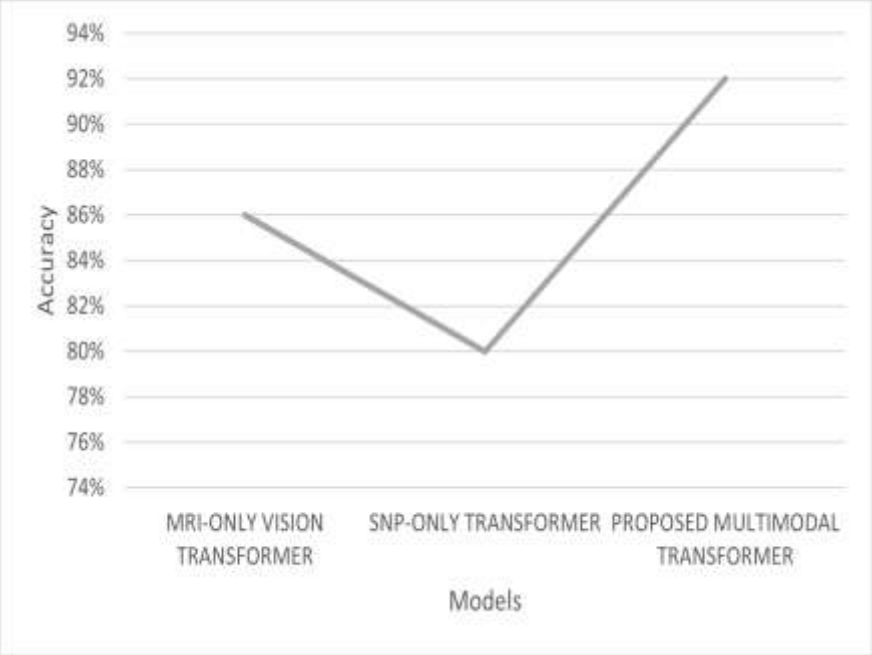

Figure 8: A comparison of Unimodal vs. Multimodal Models

Ablation Study on Fusion Strategies

To evaluate the benefit of cross-attention, we looked at a number of different fusion strategies. The output of the fusion approach is displayed in Table *6*.

Table 6: Fusion Method Performance

FUSION STRATEGY	ACCURACY
EARLY FUSION (CONCATENATION)	87%
LATE FUSION (AVERAGING EMBEDDINGS)	85%
CROSS-ATTENTION FUSION (PROPOSED)	**92%**

The results clearly show that cross-attention, which allows biologically relevant interaction between SNP variants and MRI areas, is preferable. **Figure 9** shows a graphic representation of fusion methodologies.

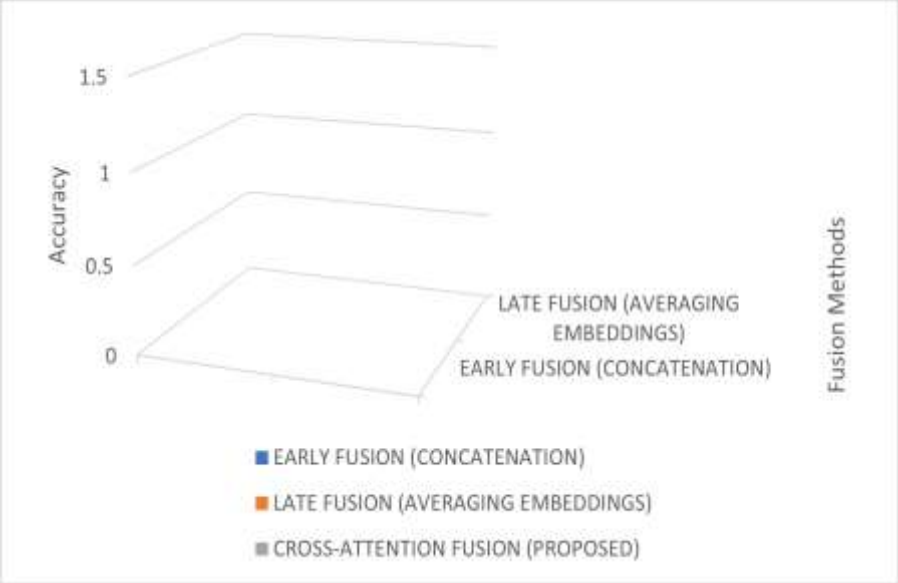

Figure 9: Fusion Method Performance

CONCLUSION

This study introduced a multimodal transformer framework that blends SNP-based genetic data with structural MRI signals to predict Alzheimer's disease. By employing bidirectional cross-attention, the model effectively learns relationships between neuroanatomical patterns and genetic risk variations, circumventing the limitations of traditional concatenation-based fusion strategies. Verified on the ADNI dataset, the proposed approach achieved 92% accuracy, outperforming both traditional fusion baselines and unimodal MRI-only and SNP-only models. The results show that a more thorough and discriminative depiction of Alzheimer's pathology can be obtained by integrating imaging and genetic modalities. Additionally, by emphasizing SNPs linked to risk and brain regions related to disease, the attention-based strategy provides interpretability. All things considered, this work demonstrates that transformer-based multimodal fusion is an actual and dependable method for primary AD detection and may help advance precision neurology.

REFERENCES

[1] G. Dolci *et al.*, "An interpretable generative multimodal neuroimaging-genomics framework for decoding Alzheimer's disease," p. arXiv: 2406.13292 v3, 2025.

[2] G. Mirabnahrazam *et al.*, "Machine learning based multimodal neuroimaging genomics dementia score for predicting future conversion to alzheimer's disease," vol. 87, no. 3, pp. 1345-1365, 2022.

[3] M. Abdelaziz, T. Wang, and A. J. F. i. a. n. Elazab, "Fusing multimodal and anatomical volumes of interest features using convolutional auto-encoder and convolutional neural networks for alzheimer's disease diagnosis," vol. 14, p. 812870, 2022.

[4] W. N. Ismail, F. Rajeena PP, and M. A. J. E. Ali, "Multforad: Multimodal mri neuroimaging for alzheimer's disease detection based on a 3d convolution model," vol. 11, no. 23, p. 3893, 2022.

[5] Q. Zuo, B. Lei, Y. Shen, Y. Liu, Z. Feng, and S. Wang, "Multimodal representations learning and adversarial hypergraph fusion for early Alzheimer's disease prediction," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 2021, pp. 479-490: Springer.

[6] J. Pan, C. Jing, Q. Zuo, M. Nieuwoudt, and S. Wang, "Cross-modal transformer GAN: a brain structure-function deep fusing framework for Alzheimer's disease," in *International Conference on Brain Inspired Cognitive Systems*, 2023, pp. 82-92: Springer.

[7] H. Guan, C. Wang, and D. J. N. Tao, "MRI-based Alzheimer's disease prediction via distilling the knowledge in multi-modal data," vol. 244, p. 118586, 2021.

[8] Q. Yu *et al.*, "A transformer-based unified multimodal framework for Alzheimer's disease assessment," vol. 180, p. 108979, 2024.

[9] E. E. Bron *et al.*, "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge," vol. 111, pp. 562-579, 2015.

[10] S. Basaia *et al.*, "Alzheimer's Disease Neuroimaging Initiative Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks," vol. 21, p. 101645, 2019.

[11] C. J. a. p. a. Zheng, "Self-Supervised Pretext Tasks for Alzheimer's Disease Classification using 3D Convolutional Neural Networks on Large-Scale Synthetic Neuroimaging Dataset," 2024.

[12] C.-F. Liu *et al.*, "Using deep Siamese neural networks for detection of brain asymmetries associated with Alzheimer's disease and mild cognitive impairment," vol. 64, pp. 190-199, 2019.

[13] J. Wen *et al.*, "Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation," vol. 63, p. 101694, 2020.

[14] J. Liu *et al.*, "Attention-Guided 3D CNN With Lesion Feature Selection for Early Alzheimer's Disease Prediction Using Longitudinal sMRI," 2024.

[15] I. E. Jansen *et al.*, "Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk," vol. 51, no. 3, pp. 404-413, 2019.

[16] J. Zhang, Y. Wang, Y. Zhang, and J. J. F. i. A. N. Yao, "Genome-wide association study in Alzheimer's disease: a bibliometric and visualization analysis," vol. 15, p. 1290657, 2023.

[17] A. GUBBINI, "Cerebellar contribution to Cognitive Impairment in early stages of Relapsing-Remitting Multiple Sclerosis: a conventional and rs-fMRI study."

[18] J.-C. Lambert *et al.*, "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease," vol. 45, no. 12, pp. 1452-1458, 2013.

[19] S. Li, K. Liu, P. J. I. T. o. C. B. Yang, and Bioinformatics, "An interpretable deep learning approach for Alzheimer's disease diagnosis using gene expression data," 2025.

[20] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, and A. s. D. N. I. J. Neuroimage, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," vol. 55, no. 3, pp. 856-867, 2011.

[21] H.-I. Suk, S.-W. Lee, D. Shen, and A. s. D. N. I. J. NeuroImage, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," vol. 101, pp. 569-582, 2014.

[22] M. K. Awang, G. Ali, and M. J. H. S. R. Faheem, "Recent Advancements in Neuroimaging-Based Alzheimer's Disease Prediction Using Deep Learning Approaches in e-Health: A Systematic Review," vol. 8, no. 5, p. e70802, 2025.

[23] J. Sheng, Y. Xin, Q. Zhang, L. Wang, and B. J. P. n. Wang, "An imaging genetics network model for clinical score assessment in Alzheimer's disease," vol. 4, no. 8, p. pgaf234, 2025.