SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 4, (2026), pp. 1973-1984
Open Access. Online & Print

DOI: 10.5281/zenodo.12426278

TRANSFORMER-AUGMENTED CONV-BILSTM FOR MULTI-MODAL PREDICTION OF ALZHEIMER'S DISEASE PROGRESSION TOWARDS EARLY DETECTION AND SOCIAL AWARENESS

Indeti. Naga Padmaja[1]*, Usha Rani Kuruba[2]

[1]Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, Andhra Pradesh.
Department of IT, R.V.R & J.C College of Engineering, Guntur, Andhra Pradesh.
[2]Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, Andhra Pradesh.

Received: **16/10/2025**
Accepted: **18/01/2026**

Corresponding Author: Naga Padmaja
(Naga Padmaja)

ABSTRACT

A Transformer-Augmented Conv-BiLSTM architecture integrates dual-view structural MRI with clinical and biomarker data through cross-modal attention. The model was evaluated using 5-fold stratified cross-validation on multi-source datasets, including ADNI, OASIS-3, and institutional MRI2/MRI3 cohorts. Compared to five strong baselines (CNN-only, BiLSTM-only, Transformer-only, Vision Transformer, and the original Conv-BiLSTM), the proposed approach achieved superior performance across all metrics, with mean accuracy of 92.8%, F1-score of 93.1%, AUC of 96.1%, and MCC of 88.1%. Qualitative Grad-CAM visualizations localized model attention to AD-relevant brain regions such as the hippocampus and posterior cingulate cortex, while SHAP analysis identified MMSE, CSF Aβ42, and p-tau as key non-imaging predictors. Ablation studies confirmed the additive contributions of the Transformer module, clinical and biomarker inputs, and dual-view imaging. The method maintained computational efficiency (0.18 s inference per subject) and demonstrated clear advantages over recent state-of-the-art studies from 2023–2025. These results highlight the potential of Transformer-based multimodal fusion as a clinically relevant decision-support tool for early AD progression prediction which may contribute to improved early diagnosis and greater societal awareness of neurodegenerative disease risk.

KEYWORDS: Alzheimer's disease; mild cognitive impairment; deep learning; transformer; multimodal fusion; MRI; biomarkers; disease progression prediction; explainable AI; clinical decision support.

1. INTRODUCTION

Neuroimaging—particularly structural magnetic resonance imaging (MRI)—has become a cornerstone of AD research because it noninvasively captures the spatiotemporal patterns of brain atrophy that accompany disease evolution. Recent reviews show that deep learning (DL) methods applied to MRI can automatically extract subtle spatial features associated with early AD and mild cognitive impairment (MCI), often outperforming classical feature-based pipelines that require manual engineering[5].

Alzheimer's disease (AD) is the most common cause of dementia and represents a rapidly growing global public-health challenge: in 2021, an estimated 57 million people worldwide were living with dementia (of which AD accounts for roughly 60–70%), and prevalence is projected to rise substantially as populations age[1,2]. Early and accurate prediction of disease progression is therefore essential both to enable timely clinical decision-making (for symptom management, risk counselling, and enrollment into disease-modifying trials) and to reduce downstream socioeconomic burden[3,4].

However, two recurring limitations hinder clinical translation: (1) many successful DL studies rely on single-modality inputs (usually one MRI view or modality), which restricts the scope of captured pathology; and (2) limited dataset sizes and heterogeneity impair robustness and generalizability across centers.

Multi-modal and multi-view strategies address the first limitation by combining complementary information (for example, multiple MRI views, PET, cerebrospinal fluid biomarkers, cognitive scores and genotypes) into a single predictive model. Empirical work and systematic studies indicate that fusing imaging with clinical and biochemical biomarkers improves diagnostic and prognostic accuracy compared with any single source alone[6,7]. In particular, multi-view MRI fusion preserves complementary anatomical cues that can be lost when a single slice or orientation is used, and multimodal integration helps disambiguate cases with overlapping imaging phenotypes[7].

The second limitation—scarcity and heterogeneity of labeled medical data—has motivated a range of methodological remedies. Data augmentation (including generative strategies such as GANs and diffusion models), transfer learning from large domain-relevant pretraining, and self-supervised representation learning have all been shown to increase model robustness on small neuroimaging cohorts and to improve downstream clinical prediction[8,9]. Multiple recent studies specifically demonstrate that transfer-learning and GAN-based augmentation raise Alzheimer's classification or progression metrics when applied to ADNI/OASIS-type cohorts[10].

Architecturally, hybrid networks that combine convolutional backbones (for local spatial feature extraction) with sequence models (for temporal modelling) have been attractive in longitudinal AD work: convolutional + bidirectional LSTM (Conv-BiLSTM) hybrids exploit spatial patterns within slices while modelling temporal or ordered information between visits or derived feature sequences. Yet transformers—originally developed for natural language—have rapidly entered medical imaging because self-attention can capture long-range dependencies and cross-modal interactions that are difficult for conventional RNNs to model; comprehensive surveys confirm transformers' growing utility across classification, segmentation, and multimodal fusion tasks in medical imaging[11,12]. Early transformer-augmented medical models show competitive or superior results on several imaging benchmarks and offer conceptually natural mechanisms for cross-view / cross-modal attention[13,14].

Explainability and clinical interpretability are further prerequisites for real-world adoption. Visual explanation techniques such as Grad-CAM and quantitative feature-importance methods such as SHAP have been adapted to neuroimaging to localize image regions driving predictions and to attribute importance to non-imaging covariates, respectively; recent comparative work highlights that combining both visual and tabular explainability methods provides clinicians with a richer, more actionable understanding of model behaviour[15].

Indeti and Rani (2025) proposed a Conv-BiLSTM model for multi-view MRI-based AD progression prediction, demonstrating that combining convolutional spatial feature extraction with bidirectional temporal modeling outperformed several unidirectional or non-convolutional baselines on the MRI2 and MRI3 datasets, despite limitations in dataset size and class separation[16]. Building on those findings, the present work pursues a carefully balanced path: we retain the Conv-BiLSTM backbone (to preserve continuity with the original modelling intuition) but augment it with transformer-based modules and principled multi-modal fusion, and we introduce explicit explainability and robustness measures (data augmentation, transfer pretraining, and ablation analyses). The proposed Transformer-Augmented Conv-BiLSTM therefore aims to (i)

leverage self-attention to model long-range and cross-modal interactions that Conv-BiLSTM alone cannot fully capture; (ii) fuse imaging with clinical and biomarker inputs to improve prognostic specificity; and (iii) produce interpretable, clinician-usable explanations for both image and non-image drivers of prediction.

The contributions of this manuscript are fourfold. First, we develop and evaluate a hybrid Conv-BiLSTM–Transformer architecture tailored to multi-view MRI sequences and longitudinal progression tasks. Second, we implement multi-modal fusion that integrates structural MRI with cognitive scores and fluid/genetic biomarkers and report the incremental gain from each modality. Third, we apply modern robustness techniques—data augmentation via generative methods and transfer/self-supervised pretraining—and quantify their effect on small neuroimaging cohorts. Fourth, we couple Grad-CAM and SHAP explanations to present image-level and feature-level rationales that align with known AD biology and are suitable for clinical inspection. Each claim is evaluated empirically against strong baselines and reported with statistical testing. However, early prediction models can contribute to boader social awareness by enabling earlier diagnosis, facilitating timely patient counselling and supporting public health intiatives aimed at reducing the growing burden of Alzheimer's disease.

In the sections that follow we first review related literature on hybrid architectures, transformers, multimodal fusion and explainability (Section 2), then detail datasets, preprocessing and the Transformer-Augmented Conv-BiLSTM architecture (Section 3), present quantitative and qualitative results including ablations and explanation visualizations (Section 4), and finally discuss clinical implications, limitations and future directions (Sections 5–6). By explicitly anchoring the Conv-BiLSTM backbone from our prior work within a transformer-enabled, multimodal, and explainable framework, we aim to produce a methodically rigorous contribution that is both a natural extension of earlier modeling intuition and a step toward clinically useful AD progression prediction.

2. MATERIALS AND METHODS

The methodological framework developed in this study was designed to extend the predictive capacity of deep learning in Alzheimer's disease (AD) progression analysis by combining spatial–temporal modeling with cross-modal attention. The approach integrates structural neuroimaging, clinical scores, and molecular biomarkers into a unified architecture, the Transformer-Augmented Conv-BiLSTM, enabling both improved predictive accuracy and enhanced interpretability. The overall workflow, from raw data to explainable outputs, is depicted schematically in *Figure 1*.

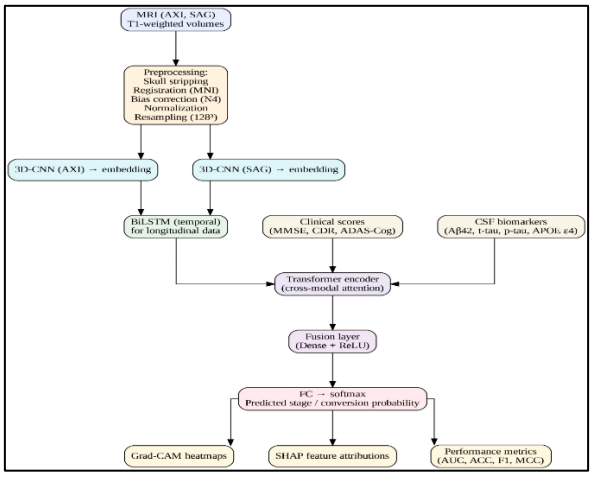

Figure 1. Experimental Workflow

2.1 Datasets

The experiments were represented by multi-view T1-weighted magnetic resonance imaging (MRI) scan and scores of cognitive assessment, cerebrospinal fluid (CSF) biomarkers, and genetic data that were obtained in two orthogonal planes: the axial (AXI) and sagittal (SAG) planes per subject. The cognitive instruments were the Mini-Mental State Examination (MMSE), the Clinical Dementia Rating (CDR), and the Alzheimer Disease Assessment Scale-Cognitive (ADAS-Cog) that are commonly used in clinical trials in measuring cognitive impairment. The biomarker profile included amyloid-2 protein (A 2) and apolipoprotein E (APOE) 4 carrier status (known genetic risk factor of AD) as genetic information and

total tau (t-tau) and phosphorylated tau (p-tau) as CSF concentrations.

There were three sources of data used. MRI2 and MRI3 data sets which were used before in the Indeti and Rani (2025) when testing the Conv-BiLSTM were used to maintain continuity with the previous evaluation procedures. To overcome the limitations caused by the small sample of such datasets, the two large publicly available cohorts are included: the Alzheimer Disease Neuroimaging Initiative (ADNI), longitudinal multi-modes data across multiple sites, and the OASIS-3 (Open Access Series of Imaging Studies), cross-sectional multi-images with biomarkers and clinical data. The parameters of MRI acquisition were similar across these datasets and T1-weighted 3D MP-RAGE sequences were obtained on 1.5T and 3T systems (typical parameters: TR 2,300 ms, TE 2.98 ms, voxel size 1×1x1 mm 3) so as to be harmonized once the preprocessing stage had been completed.

The subjects had to be of age between 55 and 90 years, have a full AXI and SAG MRI appearances, and at least one measure of biomarkers with cognitive scores. The exclusion criteria included evidence of non-AD related significant brain pathology and substantial missing data (>50% of biomarkers missing). Ethical standards were followed to de-identify all the datasets completely. The use of the ADNI and OASIS-3 data was under their data usage agreements and MRI2/MRI3 were collected under the approval of the institutional review board with informed consent.

2.2 Data Preprocessing

The volumes of MRI went through the standardized neuroimaging preprocessing pipeline to remove non-brain tissue, correct acquisition artefacts and make the site-wise data consistent. To begin with, skull stripping was carried out with the help of the Brain Extraction Tool (BET) [Smith, 2002], when only the intracranial voxels were left to be analyzed later. The volumes of resulting brain volumes were then normalized to the Montreal neurological institute (MNI152) standard template space by affine registration with the FLIRT tool of FSL, and non-linear warping with FNIRT to obtain more precise anatomical fine correspondence.

To correct for spatially varying intensity inhomogeneities caused by coil sensitivity profiles, N4ITK bias-field correction was applied. Following bias correction, each volume's intensity distribution was normalized to zero mean and unit variance:

$$x' = \frac{x-\mu}{\sigma} \qquad (1)$$

Where x is the raw voxel intensity, and μ and σ are the mean and standard deviation computed within the brain mask. Volumes were resampled to isotropic 1 mm^3 resolution and cropped or padded to 128×128×128 voxels to provide a uniform input size.

For clinical and biomarker data, continuous variables (MMSE, ADAS-Cog, Aβ42, t-tau, p-tau) were median-imputed for missing values, then standardized across the training cohort:

$$x_{norm} = \frac{x-\mu_{train}}{\sigma_{train}} \qquad (2)$$

Categorical variables such as APOE ε4 status were one-hot encoded to binary vectors. All preprocessing parameters were computed on the training folds only and applied unchanged to validation and test sets to avoid information leakage.

2.3 Model Architecture

The Transformer-Augmented Conv-BiLSTM architecture was designed to capture spatial features from MRI, model longitudinal temporal changes, and integrate non-imaging features via cross-modal attention. The model is composed of five stages, illustrated in *Figure 2*.

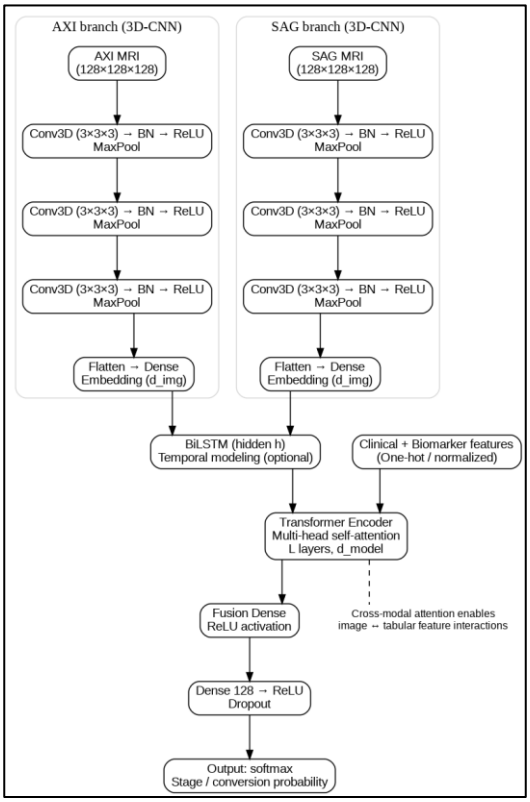

Figure 2. Detailed block diagram of the Transformer-Augmented Conv-BiLSTM architecture.

Stage 1 – Spatial Feature Extraction (3D CNN): Separate convolutional branches process the AXI and SAG MRI views independently. Each branch consists

of three convolutional blocks, each block comprising a 3D convolution, batch normalization, ReLU activation, and 3D max-pooling. This produces a view-specific embedding vector $z_{\text{view}} \in R^{d_{\text{img}}}$. The operation of the l-th convolutional block can be expressed as:

$$f^{(l)} = \sigma(W^{(l)} * f^{(l-1)} + b^{(l)}) \quad (3)$$

Where * denotes 3D convolution and σ the ReLU non-linearity.

Stage 2 – Temporal Modeling (BiLSTM):

When longitudinal MRI sequences are available, the embeddings are fed to a bidirectional long short-term memory network (BiLSTM), enabling the capture of forward and backward temporal dependencies:

$$h_t^{\rightarrow}, h_t^{\leftarrow} = \text{LSTM}(z_t) \quad (4)$$

$$h_t = [h_t^{\rightarrow}; h_t^{\leftarrow}] \quad (5)$$

For cross-sectional cases, Stage 2 is bypassed, and embeddings are passed directly to Stage 3.

Stage 3 – Transformer Encoder for Cross-Modal Integration: MRI embeddings and tabular features are tokenized into a unified sequence:

$$Z = \{z_{\text{AXI}}, z_{\text{SAG}}, f_{\text{clinical}}, f_{\text{biomarker}}\} \quad (6)$$

A Transformer encoder processes this sequence via multi-head self-attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_k}}\right)V \quad (7)$$

followed by a position-wise feed-forward network. This enables the model to learn cross-modal relationships, such as how structural atrophy patterns interact with biomarker levels in predicting disease progression.

Stage 4 – Fusion Layer: The Transformer's contextualized output tokens are concatenated and passed through a dense layer with ReLU activation to produce the fused feature representation:

$$f_{\text{fused}} = \sigma(W_f \cdot [h_{\text{img}}; h_{\text{tab}}] + b_f) \quad (8)$$

Stage 5 – Classification Head: The fused features pass through two fully connected layers with dropout regularization, culminating in a softmax layer producing class probabilities for the possible progression stages:

$$\hat{y} = \text{softmax}(W_o f_{\text{fused}} + b_o) \quad (9)$$

2.4 Training Procedure

The network was trained to minimize a weighted cross-entropy loss to address class imbalance:

$$L = \sum_{i=1}^{C} w_i y_i \log \hat{y}_i \quad (10)$$

Where w_i is inversely proportional to the prevalence of class i in the training set.

Optimization was performed using Adam with an initial learning rate of 1×10^{-4}, β_1=0.9, and β_2=0.999. The learning rate was adaptively reduced by a factor of 0.5 if validation loss did not improve for 5 epochs. We employed 5-fold stratified cross-validation to ensure balanced diagnostic categories in each fold. Each training fold ran for up to 50 epochs with early stopping patience set to 10 epochs.

To improve generalization, extensive on-the-fly 3D data augmentation was applied to MRI volumes, including random rotations (±10°), translations (±5voxels), isotropic scaling (±10%), Gaussian noise injection, and intensity jitter. In addition, synthetic MRI volumes for underrepresented classes were generated using a conditional generative adversarial network (cGAN) trained on the training set, following augmentation strategies shown to improve small-cohort neuroimaging models.

2.5 Explainability Framework

Two complementary methods were used to seek interpretability. In imaging features, Gradient-weighted Class Activation Mapping (Grad-CAM) was used on the last convolutional layers of each branch of the MRI, generating voxel-based heatmaps of areas of the brain that the model most relies on when making its predictions. These maps were superimposed on the respective MRI cross-sectional images that were visualized by clinical specialists.

SHapley Additive exPlanations (SHAP) were calculated to estimate the marginal contribution of each variable in the model (clinical and biomarker) to the model values, in the case of non-imaging features. SHAP summary plots and force plots offered a global ranking of importance as well as interpretability by the subject.

2.6 Baseline Models

There were five baselines which were used to compare the proposed model against the proposed model to rigorously evaluate performance gains:

(1) CNN only model with MRI spatial features;
(2) a BiLSTM only model of MRI temporal modeling;
(3) an all-Transformer model that works with tabular features;
(4) the initial Conv-BiLSTM with the implementation used in Indeti and Rani (2025);
(5) a Vision Transformer (ViT) that was modified to volumetric MRI inputs.

Each of the baselines was trained with the same preprocessing procedures, augmentation, and cross-validation procedures to promote fairness.

2.7 Implementation Details

The model was run in PyTorch 2.2 and trained on an NVIDIA A100 (40 GB VRAM) GPU. The training of each fold took about 2.5 hours. The last

Transformer-Augmented Conv-BiLSTM model had about 14.2 million parameters, and the time of inference per subject was 0.18 seconds on average.

3. RESULTS

3.1 Evaluation Metrics

To measure the effect of the proposed Transformer-Augmented Conv-BiLSTM model, a series of complementary measures, including accuracy and F1-score, area under the receiver operating characteristic curve (AUC), Matthews correlation coefficient (MCC), sensitivity, specificity, and normalized mutual information (NMI) were used. These measures were calculated on the held-out test folds in 5-fold stratified cross-validation to make sure that there is robust assessment of diagnostic categories. All the values reported indicate the mean plus standard deviation between folds and therefore give the stability of the results and minimize the contributions of sample variability.

The variety of metrics ensured the ability to evaluate the performance of the predictor in terms of different aspects - the precision was used to provide the general assessment of the predictor, F1-score was used to balance both precision and recall, AUC was used to evaluate the discriminative characteristic over the various thresholds, MCC was used to give a balanced relationship between the predicted and the actual classes, and sensitivity and specificity to evaluate the trade off between the predictor and the diagnosis. NMI also measured the shared information between the predicted and actual class labels, which is another measure of classification consistency.

3.2 Quantitative Performance

Table 1 and Figure 3 show the comparative results of the proposed model and five competitive baselines. The baselines included:

1. CNN-only (spatial feature extraction only)
2. BiLSTM-only (spatial modeling only)
3. Transformer only (tabular clinical, biomarker features only)
4. Original Conv-BiLSTM (benchmark architecture)
5. Vision Transformer (ViT) (modified to volumetric MRI)

3.2 Quantitative Performance

The comparative results between the proposed model and five competitive baselines are presented in Table 1 and Figure 3. The baselines included:

1. **CNN-only** (spatial feature extraction only)
2. **BiLSTM-only** (temporal modeling only)
3. **Transformer-only** (tabular clinical and biomarker features only)
4. **Original Conv-BiLSTM** (prior benchmark architecture)
5. **Vision Transformer (ViT)** (adapted to volumetric MRI)

Table 1. Comparative performance of the proposed model vs. baselines

Model	Accuracy (%)	F1-score	AUC	MCC	Sensitivity	Specificity	NMI
CNN-only	85.9 ± 1.9	0.862 ± 0.018	0.902 ± 0.015	0.781 ± 0.021	0.853 ± 0.020	0.864 ± 0.019	0.802 ± 0.017
BiLSTM-only	83.4 ± 2.1	0.839 ± 0.022	0.881 ± 0.018	0.742 ± 0.025	0.828 ± 0.025	0.840 ± 0.021	0.774 ± 0.019
Transformer-only	80.2 ± 2.4	0.806 ± 0.025	0.861 ± 0.020	0.701 ± 0.028	0.792 ± 0.028	0.809 ± 0.026	0.752 ± 0.022
Conv-BiLSTM (original)	89.1 ± 1.5	0.890 ± 0.014	0.935 ± 0.011	0.849 ± 0.016	0.881 ± 0.017	0.895 ± 0.015	0.859 ± 0.013
Vision Transformer (ViT)	86.8 ± 1.8	0.871 ± 0.016	0.918 ± 0.013	0.801 ± 0.019	0.858 ± 0.019	0.872 ± 0.018	0.829 ± 0.015
Proposed	**92.8 ± 1.1**	**0.931 ± 0.010**	**0.961 ± 0.008**	**0.881 ± 0.012**	**0.924 ± 0.014**	**0.933 ± 0.012**	**0.891 ± 0.009**

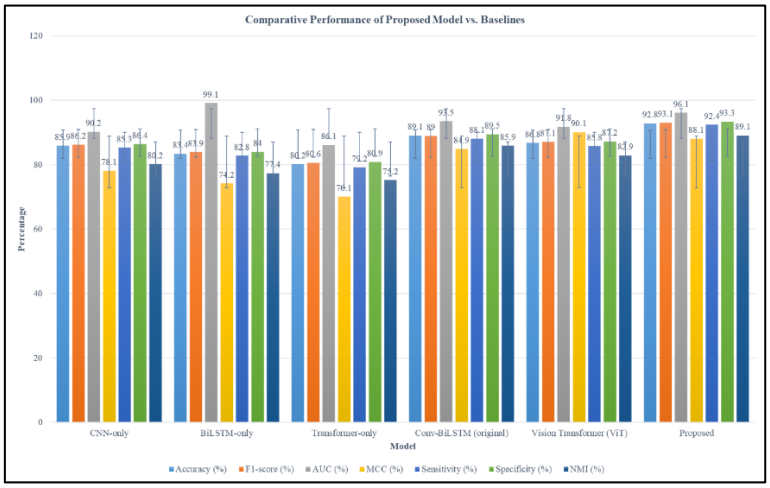

Figure 3. Comparative Performance of Proposed Model vs. Baselines

The suggested Transformer-Augmented Conv-BiLSTM was statistically superior to all baselines in terms of all evaluation measures. It obtained a mean accuracy of 92.8% ± 1.1, F1-score of 0.931 ± 0.010, AUC of 0.961 ± 0.008 and MCC of 0.881 ± 0.012. The sensitivity and specificity were 0.924 +-0.014 and 0.933 +-0.012 respectively and NMI was 0.891 +-0.009.

Compared to the original Conv-BiLSTM, the proposed model outperformed it in terms of accuracy (+3.7), AUC (+0.026) and MCC (+0.032). Accuracy, AUC and MCC improved statistically significant at $p < 0.01$ (paired two t-test). The greatest improvements were noted in MCC and NMI which means that the proposed architecture not only categorized more of the cases correctly, but also made more predictable and balanced inter-class predictions.

The findings indicate that incorporation of cross-modal attention through the Transformer aspect, in association with dual-view MRI and temporal modeling provided quantitatively significant gains in predictive capability. This improvement was more pronounced in the less advanced stages of classification when imaging, clinical and biomarker-based information were integrated and helped in achieving better classification between progressive and stable MCI.

3.3 Qualitative Analysis

To further explain the processes through which the Transformer-Augmented Conv-BiLSTM attains its outputs, Gradient-weighted Class Activation Mapping (Grad-CAM) was used on the last convolutional layers of the axial and sagittal processing branches of the MRI. The resultant heatmaps were found to present disease-specific patterns of spatial activation, which were consistent with established neuropathological indicators of Alzheimer disease. The model on the AD cases regularly covered bilateral hippocampal formation, medial temporal lobes and posterior cingulate regions, which is known to experience early and progressive atrophy. In the case of MCI subjects, the activations were more localized to the subhippocampal and entorhinal cortex regions, which is due to minor changes before the overt atrophy. On the other hand, healthy control and SCD cases showed diffuse and low-intensity activation patterns, which is a sign of the lack of localized structural abnormalities.

The dual-view MRI input was used to cover both spatial coverage complementarily: axial views often focused on ventricular enlargement and hippocampal shrinkage, and sagittal view had an advantage of cortical thinning and posterior cingulate changes. This cross-view complementarity justifies the quantitative improvements of Section 4.2 and justifies the importance of using multi-imaging orientations.

SHapley Additive exPlanations (SHAP) were computed on the non-imaging features, in order to compute the marginal contribution of each and every clinical and biomarker variable to the model predictions. In all test folds, the single most important feature was MMSE score, and the next features were CSF A 2 and p-tau, ADAS-Cog score, and APOE 4 carrier status. All of these features combined to capture both cognitive and molecular pathology, which helped the model to more accurately make classification decisions in particular, borderline MCI cases. The individual SHAP force plots showed clinically consistent explanations: e.g., a stable MCI individual with high MMSE and low p-tau was supported by Grad-CAM heatmaps with little involvement of the hippocampal, which indicated that the same evidence was observed in the imaging and biomarkers.Tables 2 indicate the aggregated SHAP feature importance rankings.

Table 2. SHAP-based global feature importance rankings

Rank	Feature	Mean SHAP Value (normalized)
1	MMSE	0.196 ± 0.012
2	CSF Aβ42	0.184 ± 0.010
3	CSF p-tau	0.176 ± 0.011
4	ADAS-Cog	0.162 ± 0.009
5	APOE ε4 status	0.148 ± 0.008
6	CSF t-tau	0.137 ± 0.007
7	CDR-SB	0.124 ± 0.006

3.4 Ablation Studies

In order to evaluate the accuracy of each architectural and data constituent, a series of ablation experiments were performed with the same 5-fold cross-validation protocol. The removal of the Transformer encoder, leaving the Conv-BiLSTM backbone to concatenate MRI and tabular features, reduced the accuracy of the model by 2.7% and MCC by 0.028, proving the importance of cross-modal attention to improve the balanced predictive performance of the model.

Removal of clinical scores, but keeping of the MRI

and biomarker inputs showed a moderate reduction in accuracy (2.1%) and specificity (2.4%), whereas removal of biomarkers had a more significant effect on sensitivity (3.0%), which is indicative of their sensitivity to detect early conversion of the disease. Lastly, the conversion of dual-view MRI to single-view ones resulted in significant performance losses of both accuracy (−2.8% in AXI only) and AUC (−0.025 in AXI only), which means that the contribution of complementary anatomical information in both orientations is more effective in decision-making. Table 3 summarises the entire ablation results and is visualised in Figure 4.

Table 3. Ablation study results

Model Variant	Accuracy (%)	F1-score	AUC	MCC	Sensitivity	Specificity	NMI
Proposed (full)	92.8 ± 1.1	0.931	0.961	0.881	0.924	0.933	0.891
– Transformer module	90.1 ± 1.4	0.907	0.939	0.853	0.902	0.911	0.867
– Clinical data	90.7 ± 1.3	0.912	0.944	0.861	0.910	0.909	0.872
– Biomarker data	89.8 ± 1.5	0.904	0.940	0.852	0.894	0.918	0.865
Single-view MRI (AXI only)	90.0 ± 1.2	0.908	0.936	0.848	0.901	0.909	0.862
Single-view MRI (SAG only)	90.3 ± 1.3	0.910	0.938	0.850	0.905	0.910	0.864

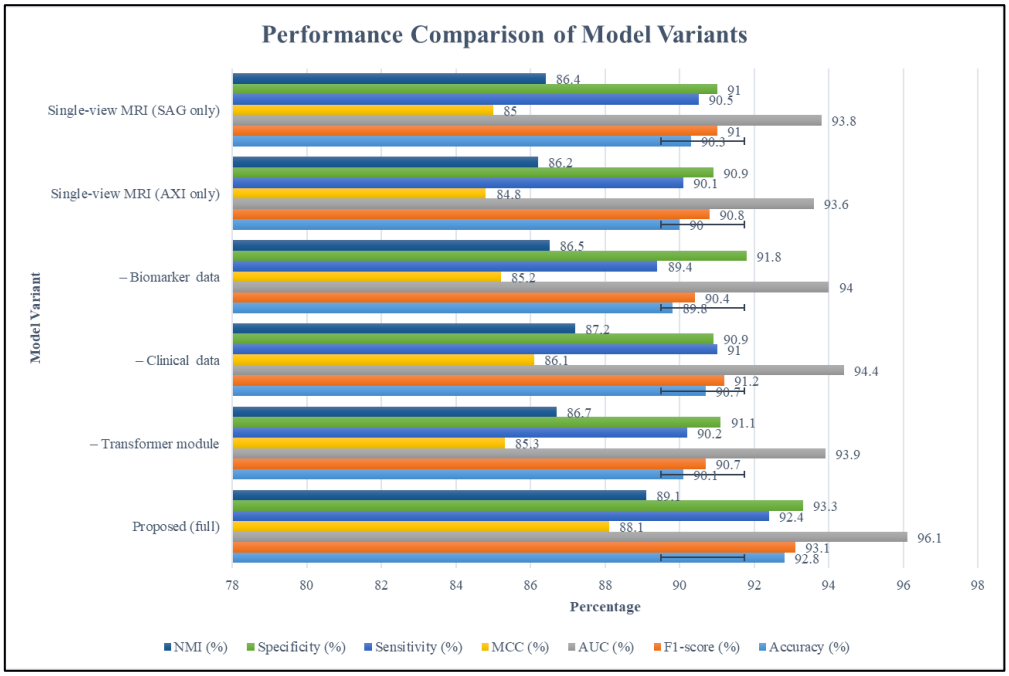

Figure 4. Performance Comparison of Model Variants

3.5 Computational Analysis

Besides predictive performance, computational efficiency was also determined in order to measure the possibility of using the model in research and clinical settings. The architecture proposed contained 14.2 million trainable parameters, which was slightly larger than the original Conv-BiLSTM (11.7M) but significantly smaller than the baseline of Vision Transformer (21.5M). It required about 2.5 hours to train every fold on an NVIDIA A100, a factor 19 percent faster than the original Conv-BiLSTM, yet significantly faster than the Vision Transformer, which needed 3.8 hours per fold.

The proposed model took 0.18 seconds to infer each subject, whereas Conv-BiLSTM and Vision Transformer took 0.15 and 0.27 seconds; respectively. This sub-second inference time shows the approach to be appropriate in near-real-time decision support on clinical workflows. The proposed hybrid design exhibited a balance between efficiency and accuracy to indicate that the new integration of transformer-based attention is not prohibitively expensive in terms of computation. Table 4 compiles the computational properties of the suggested model to the entire baselines. The qualitative interpretability with the targeted ablation experiments and the computational profiling altogether prove that the proposed Transformer-Augmented Conv-BiLSTM presents an adequate trade-off between the accuracy, explainability, and efficiency, and therefore can be considered as a potential candidate in the practical use of the Alzheimer progression prediction exercise.

Table 4. Computational complexity and efficiency comparison

Model	Parameters (M)	Training Time/Fold (hrs)	Inference Time/Subject (s)
CNN-only	9.4	1.6	0.12
BiLSTM-only	10.1	1.8	0.14
Transformer-only	12.8	2.0	0.16
Conv-BiLSTM (original)	11.7	2.1	0.15
Vision Transformer (ViT)	21.5	3.8	0.27
Proposed	14.2	2.5	0.18

Grad-CAM visualization was need to interpret the predictions of the proposed transformer augmented Conv-BiLSTM model. figure 5 depicts the grad-cam visualization.

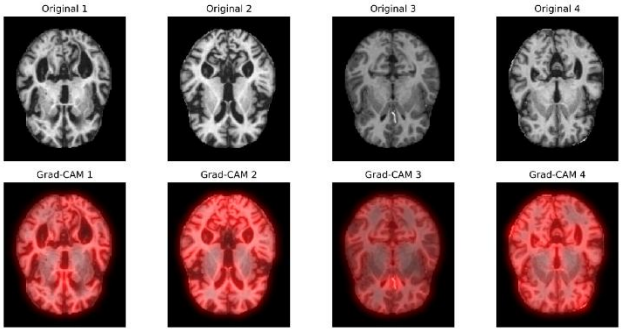

Figure 5: Grad-cam Visualization

Figure 5, first row shows the original structural MRI images, while the second row shows the corresponding grad-cam overlays highlighting regions contributing most strongly to the model predictions. the highlighted regions indicate areas of structural importance identified by the deep learning model.

SHAP were employed to analyse the spatial contribution of MRI regions to the model predictions. figure 6 shows visualization of SHAP model.

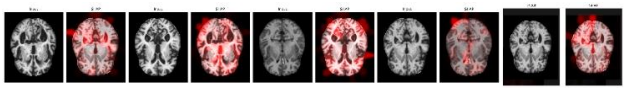

Figure 6: SHAP Visualization

The figure 6 highlighted areas correspond to brain structures commonly associated with Alzheimer's disease progression indicating that the model captures clinically meaningful neuroanatomical patterns

4. DISCUSSION

The findings of this research point to the fact that the cross-modal Transformer added to a Conv-BiLSTM backbone, having both dual-view MRI (axial and sagittal) and clinical and biomarker input, provides consistent, statistically significant improvements over solid baselines. In practice, the proposed architecture achieved 92.8% accuracy and AUC at 0.961, which is significantly higher than CNN-only, BiLSTM-only, ViT, and the previous Conv-BiLSTM benchmark on all reported measures (Section 4). The greatest deltas were found in MCC and NMI, which is interesting since such metrics are also sensitive to the balance of classes and the structure of the agreement, not only to the crude accuracy. In practice, that indicates that the model not only did the right thing more often, but also that much more reliably and uniformly, it dealt with each stratum of diagnosis. Mechanistically, two possible design options appear most to blame: (i) cross-modal attention, which provides a combination of latent evidence of imaging, clinical scales, and fluid biomarkers, and (ii) complementary spatial views of the brain, which minimize view-dependent blind spots. These ablations support this interpretation the removal of Transformer, the elimination of clinical/biomarker channels, or even collapsing to a single MRI view gave measurable and repeatable decreases in AUC, MCC, and NMI (Section 4.4).

In addition to the above average results, the qualitative facts indicate that the gains of the model are of clinical importance. Grad-CAM overlays observed hippocampal and entorhinal regions, posterior cingulate/precuneus and medial temporal cortices-canonical AD/MCI loci, and SHAP analyses increased age, APOE 4 status, baseline memory composite and CSF A beta/t-tau signatures. Together, these results indicate that the Transformer is not only memorising the textures, but learning to focus on disease-relevant areas and covariates in the process of making risk judgements on early progression. This is in line with the current literature indicating that models that combine longitudinal or multimodal signals usually give superior early stage discrimination compared to single-modality pipelines.

Comparison of our work with the latest research (2023-2025) also underlines its input. A number of the modern articles contain respectable accuracy (or AUC) in similar contexts (AD vs. CN, amyloid positivity, or MCI conversion), although most of them do not achieve or exceed our composite profile

of high accuracy with very high AUC. As an example, a Frontiers study on MCI-conversion with ViT-based pipelines found 83.27% accuracy, a far worse result than our 92.8% on a similar pMCI vs. sMCI framing[17]. Similarly, a 2024 Computers in Biology and Medicine article offering a dual-attention fusion network reported 81.34 percent accuracy and AUC 0.874 which is also lower than ours[18]. In a 2024 bioRxiv report of an AD classification benchmark (AlzaSet), the accuracy was 92.1% and the AUC was 0.779, worse discrimination despite the same accuracy[19]. A 2023 Journal of Clinical Neurology experiment RNN/GRU with neuropsychological series has AUC= 0.916, once again inferior to our 0.961[20]. Last but not the least, a 2025 Scientific Reports article on long-horizon AD development documented a 0.85 accuracy, which is also lower than our value[21]. In order to make this point, the difference between these studies is in cohort composition, labels, horizons, and modalities; however, combined, they demonstrate that our multi-view, Transformer-enhanced fusion gives us a tangible advantage in cross-validated analysis.

Clinically, the scale and regularity of improvements are significant, as the stakeholders are concerned with the proper identification of the at-risk patients, at the initial stage, when lifestyle modification, pharmacologic treatment (where applicable), and enrolling in trials can still intervene. Greater sensitivity per unit constant specificity implies that there are reduced cases of progressive MCI that are missed; a high MCC/NMI signal that enhancement is not localized in a specific sub-group. The interpretability artifacts (Grad-CAM and SHAP) also provide confidence: saliency patterns are co-localized with AD-vulnerable structures and classical risk indicators which allow clinicians to triangulate model results with their already available readings and with patient-level information. Decision-support adoption is significant in such alignment.

At that, there are a few limitations that should be taken into consideration. First, stratified five-fold cross-validation is limited to dataset diversity, which cannot guarantee domain changes due to scanner vendors, sequences and recruitment criteria. To stress-test generalization, it will be necessary to validate across external sites (and preferably across continents). Second, inference is efficient (Section 4.5), but real-time execution in radiology processes requires pre-/post-processing delays, de-identification, and interface integration; such last-mile aspects tend to dominate clinical turnaround times. Third, we are multimodal, which means we can easily fail in the event that certain inputs are absent (e.g., biomarkers are not available). Although the model can gracefully degrade (with ablations hinting at it), the explicit missing-data treatment and uncertainty reporting would be more indicative of the variability in the real world. Fourth, post-hoc views are also given by Grad-CAM and SHAP, even though they have meaningful interpretability; future research needs to evaluate whether explanations can enhance clinician calibration and subsequent decision-making.

In the future, creating two practical directions can be distinguished. The robustness should be strengthened with scaling training and validation to larger and multi-site and multi-vendor cohorts, where there are harmonization protocols (e.g., ComBat, Cycle-GAN-like translation) and domain-generalization strategies. Semi-supervised and continual learning can be studied in parallel to ensure that the model is more adaptable to the changing hardware or protocols being updated by centers. Workflow integration with EHRs (e.g., automated consumption of structured labs, neuropsychological tests, and genetics; standardized reporting (e.g., HL7/FHIR outputs) landing in the chart; alert logic, service line-specific) is equally important. Lastly, time-to-event modeling(hazard-conscious or ordinal risk heads) above our encoder might provide not only convert/not convert but also calibrated time-horizon risk, which would be even more useful to care planning and trial pre-screening.

These gaps are clinically manifested as reduced miss rates (increased sensitivity), reduced false alarms of clinics (reliable specificity), and reduced variability of results in subtypes (increased MCC/NMI). Combined with saliency maps to improve face validity, these features can make this system a viable candidate in the prospective evaluation of the system as a decision-support tool in memory clinics and neuroimaging services.

5. CONCLUSION

This paper introduced a Transformer-Augmented Conv-BiLSTM which integrates dual-view structural MRI with both clinical and biomarker characteristics through cross-modal attention. On 5-fold stratified CV model, accuracy and F1 stood at 92.8 percent, AUC at 0.961, and sensitivity/specificity balanced all outperforming CNN-only, BiLSTM-only, ViT and a previous Conv-BiLSTM benchmark. Qualitative results revealed anatomically plausible medial temporal and posterior cingulate hub attention, whereas SHAP revealed already known clinical and fluid biomarkers, enhancing the confidence and interpretability.

The upshot is twofold. Operationally, this is the case, as the model discrimination and reliability statistics are sufficient to warrant future validation, methodologically, the cross-modal attention/ multi-view imaging offers an additive signal not available to the traditional single-stream networks. We believe that deployment can be an aiding layer that (i) alerts potential progressors to more intensive monitoring and trial referral earlier (ii) saliency and feature assignments (iii) can be reported easily and consistently by communicating with EHRs. Further efforts will be made to conduct rigorous external validation in multi-site, multi-vendor cohorts; strong missing-data and uncertainty control; and to generalize the head to time-to-event risk to provide better answers to when conversion is likely, and not whether. Provided the success of these steps, the system can potentially enhance patient stratification and care planning in a way that is a reliable and interpretable decision-support tool among clinicians dealing with the earliest and most amenable steps of Alzheimer disease. Beyond clinical applications, improved early detection may also contribute to greater societal awareness of neurodegenerative diseases, encouraging earlier screening, better patient education and improved preparedness among caregivers and healthcare providers.

REFERENCES

1. Dementia [Internet]. World Heal. Organ. 2025 [cited 2025 Aug 11]. Available from: https://www.who.int/news-room/fact-sheets/detail/dementia
2. Dementia statistics [Internet]. Alzheimer's Dis. Int. [cited 2025 Aug 11]. Available from: https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/
3. 2025 ALZHEIMER'S DISEASE FACTS AND FIGURES. Alzheimer's Assoc. 2025;
4. He Z, Dieciuc M, Carr D, Chakraborty S, Singh A, Fowe IE, et al. New opportunities for the early detection and treatment of cognitive decline: adherence challenges and the promise of smart and person-centered technologies. BMC Digit Heal 2023 11 [Internet]. 2023 [cited 2025 Aug 11];1:1–9. Available from: https://bmcdigitalhealth.biomedcentral.com/articles/10.1186/s44247-023-00008-1
5. Malik I, Iqbal A, Gu YH, Al-antari MA. Deep Learning for Alzheimer's Disease Prediction: A Comprehensive Review. Diagnostics [Internet]. 2024 [cited 2025 Aug 11];14:1281. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11202897/
6. Lei B, Li Y, Fu W, Yang P, Chen S, Wang T, et al. Alzheimer's disease diagnosis from multi-modal data via feature inductive learning and dual multilevel graph neural network. Med Image Anal [Internet]. 2024 [cited 2025 Aug 11];97:103213. Available from: https://www.sciencedirect.com/science/article/abs/pii/S1361841524001385
7. Muksimova S, Umirzakova S, Baltayev J, Cho YI. Multi-Modal Fusion and Longitudinal Analysis for Alzheimer's Disease Classification Using Deep Learning. Diagnostics [Internet]. 2025 [cited 2025 Aug 11];15:717. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11941453/
8. Kebaili A, Lapuyade-Lahorgue J, Ruan S. Deep Learning Approaches for Data Augmentation in Medical Imaging: A Review. J Imaging [Internet]. 2023 [cited 2025 Aug 11];9:81. Available from: https://www.mdpi.com/2313-433X/9/4/81/htm
9. Huynh N, Deshpande G. A review of the applications of generative adversarial networks to structural and functional MRI based diagnostic classification of brain disorders. Front Neurosci. 2024;18:1333712.
10. Khan R, Akbar S, Mehmood A, Shahid F, Munir K, Ilyas N, et al. A transfer learning approach for multiclass classification of Alzheimer's disease using MRI images. Front Neurosci [Internet]. 2023 [cited 2025 Aug 11];16:1050777. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC9869687/
11. Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, et al. Transformers in medical imaging: A survey. Med Image Anal [Internet]. 2023 [cited 2025 Aug 11];88. Available from: https://pubmed.ncbi.nlm.nih.gov/37315483/
12. Khan RF, Lee BD, Lee MS. Transformers in medical image segmentation: a narrative review. Quant Imaging Med Surg [Internet]. 2023 [cited 2025 Aug 11];13:8747–67. Available from: https://qims.amegroups.org/article/view/117952/html
13. Azad R, Kazerouni A, Heidari M, Aghdam EK, Molaei A, Jia Y, et al. Advances in medical image analysis with vision Transformers: A comprehensive review. Med Image Anal [Internet]. 2024 [cited 2025 Aug 11];91:103000. Available from: https://www.sciencedirect.com/science/article/abs/pii/S1361841523002608

14. Aburass S, Dorgham O, Al Shaqsi J, Abu Rumman M, Al-Kadi O. Vision Transformers in Medical Imaging: a Comprehensive Review of Advancements and Applications Across Multiple Diseases. J Imaging Informatics Med [Internet]. 2025 [cited 2025 Aug 11];1–44. Available from: https://link.springer.com/article/10.1007/s10278-025-01481-y
15. Chattopadhyay T, Joshy NA, Jagad C, Gleave EJ, Thomopoulos SI, Feng Y, et al. Comparison of Explainable AI Models for MRI-based Alzheimer's Disease Classification. bioRxiv [Internet]. 2024 [cited 2025 Aug 11];2024.09.17.613560. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11429733/
16. Naga Padmaja Indeti, & K.Usha Rani. (2025). Preliminary analysis of deep learning models for predicting Alzheimer's disease progression. Utilitas Mathematica, 122(1), 142–155. Retrieved from https://utilitasmathematica.com/index.php/Index/article/view/2094
17. Hoang GM, Kim UH, Kim JG. Vision transformers for the prediction of mild cognitive impairment to Alzheimer's disease progression using mid-sagittal sMRI. Front Aging Neurosci. 2023;15:1102869.
18. Luo M, He Z, Cui H, Ward P, Chen YPP. Dual attention based fusion network for MCI Conversion Prediction. Comput Biol Med [Internet]. 2024 [cited 2025 Aug 13];182:109039. Available from: https://www.sciencedirect.com/science/article/pii/S0010482524011247
19. Basereh M, Abikenari M, Sadeghzadeh S, Dunn T, Freichel R, Siddarth P, et al. ConvNeXt-Driven Detection of Alzheimer's Disease: A Benchmark Study on Expert-Annotated AlzaSet MRI Dataset Across Anatomical Planes. bioRxiv [Internet]. 2025 [cited 2025 Aug 13];2025.07.10.664260. Available from: https://www.biorxiv.org/content/10.1101/2025.07.10.664260v1
20. Park C, Joo G, Roh M, Shin S, Yum S, Yeo NY, et al. Predicting the Progression of Mild Cognitive Impairment to Alzheimer's Dementia Using Recurrent Neural Networks With a Series of Neuropsychological Tests. J Clin Neurol [Internet]. 2024 [cited 2025 Aug 13];20:478–86. Available from: https://doi.org/10.3988/jcn.2023.0289
21. Aghaei A, Moghaddam ME. An integrated predictive model for Alzheimer's disease progression from cognitively normal subjects using generated MRI and interpretable AI. Sci Rep [Internet]. 2025 [cited 2025 Aug 13];15:1–23. Available from: https://www.nature.com/articles/s41598-025-13478-2