International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2s) 2026 pp. 1255-1268
https://doi.org/10.70917/ijcisim-2026-2040
Article

HYBRID TRANSFORMER-ENHANCED 3D CNN FOR EXPLAINABLE ALZHEIMER'S DISEASE CLASSIFICATION AND PROGRESSION PREDICTION

Vinutha H.[1]*, Kavita V. Horadi[2], Pavan G. Malghan[3], Asha M. S.[4], J. Gul Shaira Banu[5], Anoop G. L.[6]

[1] Department of Machine Learning (AI & ML), B.M.S. College of Engineering (BMSCE), Bengaluru, India
[2] Department of Computer Science and Engineering, BNMIT, VTU, India
Email: kavita.bnmit3@gmail.com
[3] BMS Institute of Technology & Management (BMSIT&M), Yelahanka, Bengaluru, India
Email: pavan.m@bmsit.in
[4] Department of Computer Science and Engineering, CHRIST University, Bengaluru, India
Email: ashamsgowda05@gmail.com
[5] Department of Computer Science and Engineering, PMC Tech, Hosur, India
Email: drgulshairabanu@gmail.com
[6]Department of Computer Science and Engineering, CHRIST University, Bengaluru, Karnataka, India
Email: gl.anoop1@gmail.com
Corresponding Author: Vinutha H. (Email: vinuthah.mel@bmsce.ac.in)

Abstract: Alzheimer's disease (AD) is an irreversible neurodegenerative disorder whose timely and interpretable diagnosis remains a major clinical challenge. Three-dimensional convolutional neural networks (3D CNNs) capture fine-grained local morphometric changes from structural magnetic resonance imaging (sMRI), yet their limited receptive field hinders modelling of long-range anatomical dependencies, while pure transformer models are data-hungry and difficult to interpret. This paper proposes a hybrid Transformer-Enhanced 3D CNN (TE-3DCNN) that couples a squeeze-and-excitation residual 3D CNN encoder with a transformer encoder through a gated cross-module fusion mechanism, and augments it with a dedicated progression-prediction head for mild cognitive impairment (MCI)-to-AD conversion. To make predictions trustworthy, the framework integrates Grad-CAM++ and self-attention rollout to generate volumetric saliency maps. Experiments on the public ADNI sMRI cohort for three-way classification (cognitively normal, MCI, AD) show that TE-3DCNN attains 94.8% accuracy, 94.7% macro F1-score and 0.979 mean AUC, outperforming a 3D ResNet, a 3D Vision Transformer and a convolution–Swin transformer baseline by 2.7–6.4 percentage points. Ablation studies confirm the complementary contribution of the CNN encoder, transformer branch, gated fusion and squeeze-and-excitation recalibration. The produced saliency maps consistently localise the hippocampus, medial temporal lobe and ventricular regions, agreeing with established AD neuropathology. The results indicate that TE-3DCNN offers an accurate, explainable and clinically meaningful tool for AD diagnosis and prognosis.

Keywords: Alzheimer's disease, 3D convolutional neural network, vision transformer, explainable AI, Grad-CAM++, progression prediction, structural MRI, hybrid deep learning

1. INTRODUCTION

Alzheimer's disease (AD) is the most prevalent form of dementia, accounting for an estimated 60–70% of the more than 55 million dementia cases worldwide, and its incidence is projected to triple by 2050 as the global population ages [1]. AD is characterised by the progressive accumulation of amyloid-β plaques and neurofibrillary tau tangles, which precipitate synaptic loss and macroscopic neurodegeneration, most notably hippocampal and entorhinal atrophy together with ventricular enlargement. Because the underlying neuronal damage is irreversible, the greatest clinical benefit is obtained when the disease, or its prodromal stage of mild cognitive impairment (MCI), is identified early enough for therapeutic and lifestyle interventions to slow decline [2, 3]. Structural magnetic resonance imaging (sMRI) is widely used for this purpose because it is non-invasive, broadly available and sensitive to the morphometric signatures of neurodegeneration. Manual radiological assessment, however, is time-consuming, subjective and unable to detect the subtle, spatially distributed changes that distinguish stable MCI from progressive MCI.

Deep learning has transformed automated AD diagnosis from sMRI. Three-dimensional convolutional neural networks (3D CNNs) operate directly on volumetric data and learn hierarchical, translation-equivariant features that encode local cortical and subcortical morphology [4, 5, 6]. Despite their success, the intrinsically local receptive field of convolution makes it difficult for 3D CNNs to model the long-range dependencies between distant but functionally coupled brain regions that jointly characterise AD. Vision transformers (ViTs) address this limitation through global self-attention [7, 8], but they lack the inductive biases of convolution, require large annotated datasets that are scarce in neuroimaging, and are computationally demanding when applied to high-resolution 3D volumes. Hybrid CNN–transformer architectures, which combine the local feature-extraction strength of convolution with the global context modelling of attention, have therefore emerged as a promising direction across medical imaging [9, 10, 11].

Two further requirements limit clinical translation. First, accurate classification of the current diagnostic stage is insufficient; clinicians need a reliable estimate of whether and when an MCI patient will convert to AD, i.e. progression prediction [2, 12]. Second, deep models are frequently treated as black boxes, which undermines clinical trust; explainable artificial intelligence (XAI) techniques such as Grad-CAM and attention visualisation are needed to expose the anatomical evidence behind each decision [13, 14, 15]. Most existing studies tackle these objectives in isolation, optimising classification accuracy while neglecting interpretability or prognosis.

To bridge these gaps, this work proposes a unified, explainable Hybrid Transformer-Enhanced 3D CNN (TE-3DCNN). A squeeze-and-excitation (SE) residual 3D CNN encoder extracts discriminative local descriptors, which are tokenised and refined by a transformer encoder that captures global inter-regional context. A gated cross-module fusion block adaptively balances the two representations, after which a classification head predicts the diagnostic label and a progression head estimates MCI-to-AD conversion risk. Grad-CAM++ and self-attention rollout jointly produce volumetric saliency maps that localise the evidence. Figure 1 presents the overall framework. The main contributions of this paper are summarised as follows:

A novel hybrid TE-3DCNN architecture is proposed that couples an SE-residual 3D CNN with a transformer encoder through a learnable gated fusion mechanism, jointly modelling local morphometry and global anatomical context from sMRI.

A multi-task design is introduced that performs three-way diagnostic classification (CN/MCI/AD) and MCI-to-AD progression prediction within a single end-to-end framework.

An integrated explainability module combines Grad-CAM++ with self-attention rollout to produce clinically interpretable 3D saliency maps that are quantitatively and qualitatively validated against known AD pathology.

Extensive experiments on the public ADNI cohort, including comparison with three recent architectures and a comprehensive ablation study, demonstrate state-of-the-art accuracy together with transparent, trustworthy predictions.

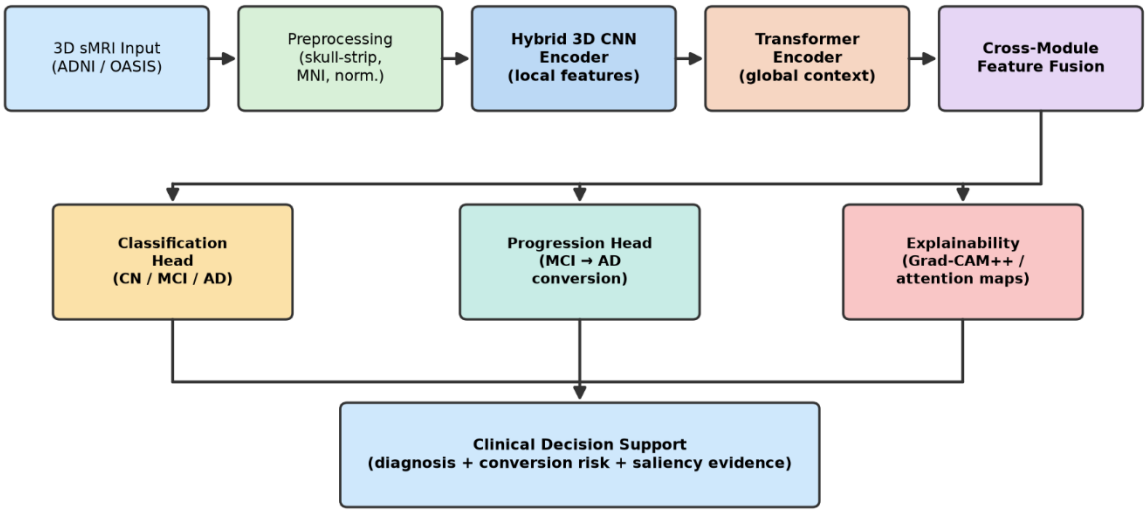

Fig. 1. Proposed explainable Hybrid Transformer-Enhanced 3D CNN (TE-3DCNN) framework for Alzheimer's disease classification and progression prediction.

The remainder of this paper is organised as follows. Section II reviews related work on deep learning for AD diagnosis, hybrid CNN–transformer models, progression prediction and explainability. Section III details the proposed methodology, including the network architecture, mathematical formulation and training algorithm. Section IV describes the dataset, experimental setup and results, including comparison and ablation studies. Section V discusses the findings and limitations, and Section VI concludes the paper.

2. RELATED WORK

A. 3D CNN-based Alzheimer's Disease Diagnosis

Convolutional networks remain the backbone of sMRI-based AD analysis. Feng et al. [4] proposed a 3D multi-feature fusion convolutional network that aggregates complementary volumetric descriptors to improve diagnostic discrimination. Zarei et al. [5] combined a 3D CNN with hand-crafted radiomic features from T1-weighted MRI for CN/MCI/AD classification, reporting that learned and radiomic representations are complementary. Contrary to the prevailing trend toward ever-deeper models, Grødem et al. [6] showed that extremely small CNNs can classify AD competitively, highlighting the risk of over-parameterisation on limited neuroimaging cohorts. Transfer learning has also been widely adopted; Ghaffari et al. [16] fine-tuned pretrained deep networks for fully automated AD detection, mitigating data scarcity. Although these methods capture local atrophy patterns effectively, their restricted receptive fields limit the modelling of distributed, whole-brain alterations.

B. Transformers and Hybrid Architectures in Medical Imaging

Transformers have rapidly permeated medical image analysis, as surveyed comprehensively by Shamshad et al. [17]. For dementia, Huang and Qiu [7] introduced an ensemble vision transformer that improves robustness through model diversity, while Alp et al. [8] designed a joint transformer architecture for 3D MRI AD classification. Because pure attention discards convolutional inductive biases, hybrid designs have gained traction. Zhao et al. [9] equipped a CNN with a vision transformer for 3D MRI AD diagnosis, and Khatri and Kwon [18] developed an optimised lightweight convolution–attention model for structural MRI. Hu et al. [10] proposed Conv-Swinformer, integrating convolution with shifted-window attention for AD classification, which we adopt as a strong baseline. Beyond AD, hybrid transformers have advanced brain tumour segmentation through clinical-knowledge-driven cross-attention [11], cross-dataset brain-tissue segmentation [19] and high-resolution MRI synthesis [20], confirming the generality of coupling convolution with attention for volumetric neuroimaging.

C. Multimodal and Progression Modelling

Because no single modality fully characterises AD, multimodal fusion has been extensively studied. Tang et al. [21] proposed CsAGP, a dual-transformer with cross-attention and graph pooling for multimodal AD detection, and

Abdelaziz et al. [22] introduced a multi-scale multimodal framework integrating imaging and genetic markers. Gao et al. [23] addressed the common problem of missing modalities with a multimodal transformer that generates incomplete images while diagnosing AD, and Fedorov et al. [24] employed self-supervised multimodal learning to discover disorder-relevant brain regions without labels. Progression prediction has been tackled by Al Olaimat et al. [2], whose PPAD architecture forecasts AD progression from longitudinal data, and by Baytas [12], who studied the robustness of MCI-to-AD prediction under adversarial perturbation. Yue et al. [3] further demonstrated that deep learning can reveal subtle pre-MCI changes predictive of future cognitive decline. These studies motivate the integration of a dedicated progression head within our framework. Beyond architecture and modality, the training paradigm also shapes real-world deployment: because neuroimaging data are siloed across institutions and governed by strict privacy regulations, communication-efficient federated learning (CEFL) has been proposed to classify medical images collaboratively across bandwidth-constrained healthcare networks without centralising patient data [25], with hybrid CEFL schemes that combine gradient sparsification and quantization further reducing communication overhead [26]. Such privacy-preserving paradigms are complementary to the centralised model proposed here and motivate the multi-site extensions discussed in Section V.

D. Explainable AI for Neuroimaging

Interpretability is indispensable for clinical adoption. Khosroshahi et al. [14] surveyed XAI in AD neuroimaging and emphasised the need for anatomically faithful explanations, while Mahmud et al. [13] combined deep transfer learning with explainability for AD diagnosis. Vetrithangam et al. [15] fused a deep CNN with enhanced weighted fuzzy c-means clustering to produce interpretable detections. Comprehensive reviews by Khojaste-Sarakhsi et al. [27] and Kale et al. [1] conclude that the field is moving toward integrated systems that are simultaneously accurate, interpretable and capable of prognosis. To the best of our knowledge, however, few prior works unify hybrid CNN–transformer representation learning, multi-task classification and progression prediction, and built-in volumetric explainability within a single end-to-end model - the gap addressed by the proposed TE-3DCNN.

3. PROPOSED METHODOLOGY

The proposed TE-3DCNN is an end-to-end, multi-task framework comprising five components: (i) an image preprocessing pipeline, (ii) an SE-residual 3D CNN encoder, (iii) a transformer encoder, (iv) a gated cross-module fusion block feeding classification and progression heads, and (v) an explainability module. Figure 2 shows the detailed architecture and Figure 3 details the explainability and progression sub-modules.

A. Image Preprocessing

All volumes undergo a standard pipeline: non-brain tissue removal (skull stripping), affine registration to the MNI-152 template to achieve spatial correspondence across subjects, bias-field correction and resampling to an isotropic 1.5 mm grid of size 113×137×113. Voxel intensities within the brain mask are standardised by region-of-interest z-score normalisation to suppress scanner-induced variability, as defined in (1):

$$x' = \frac{x - \mu_{\mathrm{ROI}}}{\sigma_{\mathrm{ROI}} + \varepsilon} \qquad (1)$$

where x and x′ denote the original and normalised intensities, μ_{ROI} and σ_{ROI} are the mean and standard deviation within the brain mask, and ε is a small constant ensuring numerical stability. Online augmentation (random flips, ±10° rotations, elastic deformations and intensity jitter) is applied during training.

B. SE-Residual 3D CNN Encoder

The encoder consists of a stem followed by four residual stages with output widths of 32, 64, 128 and 256 channels, each performing volumetric feature extraction. A 3D convolution at layer l is expressed as (2):

$$F^{(l)} = \varphi\, W^{(l)} \circledast F^{(l-1)} + b^{(l)} \qquad (2)$$

where $\circledast$ denotes 3D convolution, W^l and b^l are the learnable kernels and biases, F^{l-1} is the input feature volume, and φ(·) is the GELU activation with group normalisation. To emphasise informative channels, each residual block embeds a squeeze-and-excitation (SE) unit that recalibrates feature maps as in (3):

$$\mathrm{F} = \sigma(W_2\delta\big(W_1 GAP(F)\big) \odot F \qquad (3)$$

where GAP(·) is global average pooling over the spatial dimensions, W_1 and W_2 are the bottleneck projection weights, δ(·) and σ(·) denote the ReLU and sigmoid functions, and $\odot$ is channel-wise multiplication. The recalibrated

output is added to the block input through a residual shortcut, stabilising optimisation of the deep volumetric network. The final encoder produces a feature volume that is globally average-pooled into a compact descriptor f_{cnn} and, in parallel, forwarded to the transformer branch.

C. Transformer Encoder

The last-stage feature volume is partitioned into a sequence of non-overlapping 3D patches that are linearly projected into D-dimensional tokens. A learnable class token x_{cls} is prepended and learnable 3D positional embeddings E_{pos} are added to retain spatial order, forming the input sequence (4):

$$z_0 = [x_{\text{cls}};\ t_1;\ \dots\ ;\ t_N] + E_{\text{pos}} \quad (4)$$

The sequence is processed by L transformer layers. The core operation is scaled dot-product self-attention (5), where the queries, keys and values are linear projections $Q = zW_Q, K = zW_K$ and $V = zW_V$:

$$\text{Attention}(Q,K,V) = Softmax\left(\frac{QK^{\intercal}}{\sqrt{d_k}}\right)V \quad (5)$$

where d_k is the key dimension that scales the dot products. Multi-head self-attention (MHSA) concatenates h parallel attention heads to jointly attend to different representation subspaces, as in (6):

$$\text{MHSA}\, z = \text{Concat}(\text{head}_1, \dots, \text{head}_h)\, W_O \quad (6)$$

where $head_i = Attention(zW_i^Q, zW_i^K, zW_i^V)$ and W_O is the output projection. Each transformer layer applies MHSA and a position-wise feed-forward network (MLP), each wrapped by layer normalisation (LN) and a residual connection, as in (7) and (8):

$$z'_l = MHSA\big(LN(z_{l-1})\big) + z_{l-1} \quad (7)$$

$$z_l = MLP(LN(z'_l) + z'_l \quad (8)$$

The transformer output corresponding to the class token, $f_{cls} = z_L^{(0)}$, encodes global inter-regional context that complements the local descriptor f_{cnn} .

D. Gated Cross-Module Fusion

Rather than naively concatenating the two representations, a gated fusion mechanism learns to weight the local and global features adaptively. A gate vector g is computed in (9) and used to combine the descriptors in (10):

$$g = \sigma(W_g\,[f_{\text{cnn}};\ f_{\text{cls}}] + b_g) \quad (9)$$

$$f_{\text{fused}} = g \odot f_{\text{cnn}} + (1 - g) \odot f_{\text{cls}} \quad (10)$$

where W_g and b_g are learnable parameters and $[\cdot;\cdot]$ denotes concatenation. The fused representation f_{fused} is shared by the two task heads.

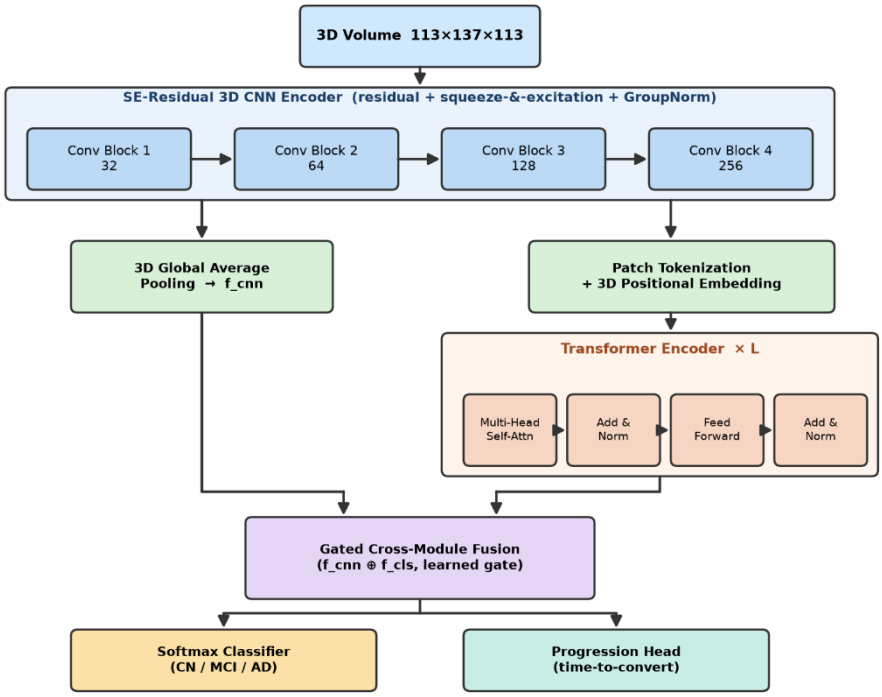

Fig. 2. Detailed architecture of the proposed TE-3DCNN, showing the SE-residual 3D CNN encoder, feature tokenisation, the L-layer transformer encoder, gated cross-module fusion, and the classification and progression heads.

E. Classification and Progression Heads

The classification head maps f_{fused} to the three diagnostic classes through a fully connected layer and a softmax, giving the posterior probability of class k in (11):

$$p(y = k \mid x) = \frac{\exp(o_k)}{\sum_{j=1}^{C} \exp o_j} \qquad (11)$$

where o_k is the logit of class k and C = 3. The progression head models MCI-to-AD conversion as a time-to-event problem; a Cox proportional-hazards layer on f_{fused} estimates the instantaneous hazard in (12):

$$h(t \mid x) = h_0(t)\ \exp(\beta^{\top} f_{\text{fused}}) \quad (12)$$

where $h_0(t)$ is the baseline hazard and β are the risk coefficients. The conversion probability over a clinical horizon is obtained from the corresponding survival function, enabling individualised risk stratification (Figure 3(b)).

F. Explainability Module

To expose the anatomical evidence behind each prediction, Grad-CAM++ is applied to the final convolutional feature maps. The importance weight of channel k for target class c is obtained from the back-propagated gradients, and the class-discriminative localisation map is computed in (13):

$$L_{\text{CAM}}^{c} = \text{ReLU}(\sum_k \alpha_k^c A^k) \qquad (13)$$

where A^k is the k-th activation map and α_k^c are the Grad-CAM++ weights derived from first-, second- and third-order gradients of the class score. The convolutional saliency is fused with self-attention rollout from the transformer branch to yield a unified 3D heat map, providing both local and global interpretability (Figure 3(a)).

G. Training Objective and Algorithm

The framework is optimised end-to-end with a composite objective (14) that jointly supervises classification and progression while regularising the parameters θ:

$$L = L_{\text{cls}} + \lambda_1 L_{\text{prog}} + \lambda_2 \|\theta\|^2 \quad (14)$$

where L_{cls} is the categorical cross-entropy classification loss, L_{prog} is the negative partial log-likelihood of the Cox model, and λ_1, λ_2 are weighting hyper-parameters. The complete training procedure is summarised in Algorithm 1.

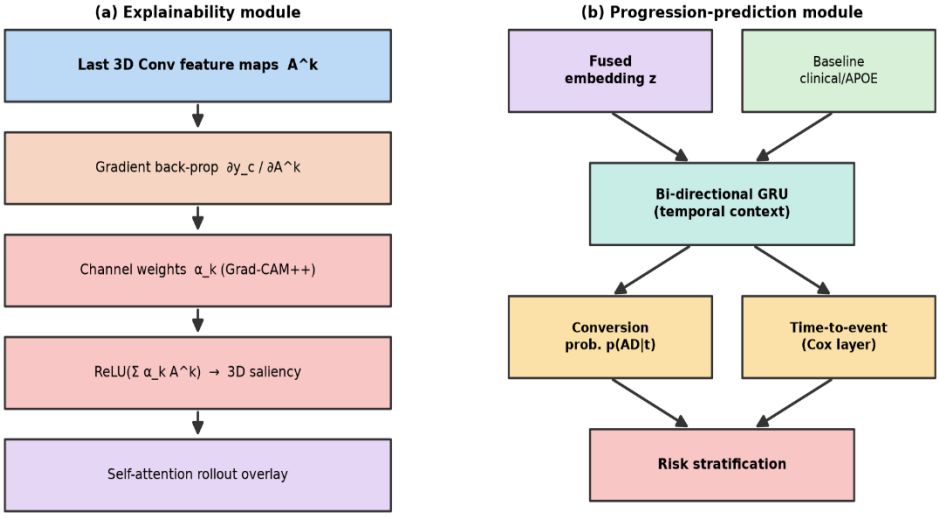

Fig. 3. (a) Explainability module combining Grad-CAM++ saliency with self-attention rollout; (b) progression-prediction module fusing the learned embedding with baseline covariates through a recurrent temporal encoder and a Cox time-to-event layer.

Algorithm 1: Training of the Hybrid Transformer-Enhanced 3D CNN (TE-3DCNN)

Input: preprocessed 3D sMRI volumes $\{X_i\}$, diagnostic labels $\{y_i\}$,
conversion times/events $\{t_i, e_i\}$; hyper-parameters $\lambda 1$, $\lambda 2$, η

Output: trained parameters θ of the TE-3DCNN

```
Initialise encoder, transformer, fusion and heads; θ ← θ0
for epoch = 1 to E do
    for each mini-batch B do
        X ← Augment(Normalise(B))                    // Eq. (1)
        F ← SE_Residual_3DCNN_Encoder(X)             // Eq. (2)-(3)
        f_cnn ← GAP(F);  Z ← Tokenise(F) + E_pos      // Eq. (4)
        f_cls ← Transformer_Encoder(Z)               // Eq. (5)-(8)
        f_fused ← GatedFusion(f_cnn, f_cls)          // Eq. (9)-(10)
        p ← Softmax_Classifier(f_fused)              // Eq. (11)
        h ← Cox_Progression_Head(f_fused)            // Eq. (12)
        L ← L_cls(p,y) + λ1·L_prog(h,t,e) + λ2·||θ||^2  // Eq. (14)
        θ ← θ − η · ∇θ L                 // AdamW update
    end for
    Evaluate on validation set; apply early stopping
```

```
    15: end for
    16: Generate Grad-CAM++ and attention-rollout saliency  // Eq. (13)
    17: return θ
```

4. EXPERIMENTS AND RESULTS

A. Dataset Details

Experiments were conducted on the publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu), an open-access, multi-site longitudinal study that provides T1-weighted structural MRI together with clinical and genetic metadata [2, 17]. A balanced cohort of 1.5T/3T T1-weighted scans was assembled across three diagnostic groups - cognitively normal (CN), mild cognitive impairment (MCI) and Alzheimer's disease (AD). To prevent data leakage, the dataset was partitioned at the subject level into training, validation and test sets in a 70:10:20 ratio, ensuring that all scans from a given subject reside in a single split. Table I summarises the composition. For progression analysis, the MCI cohort was further labelled as stable MCI (sMCI) or progressive MCI (pMCI) using a 36-month follow-up window. All volumes were preprocessed as described in Section III-A. Figure 4 shows representative preprocessed axial slices for each class, in which the progressive ventricular enlargement and cortical atrophy characteristic of AD are clearly visible.

TABLE I. Composition of the ADNI T1-weighted MRI dataset (subject-level split).

Diagnostic class	Subjects	Scans	Train	Validation	Test
Cognitively Normal (CN)	512	1500	1050	150	300
Mild Cognitive Impairment (MCI)	468	1360	952	136	272
Alzheimer's Disease (AD)	372	1080	756	108	216
Total	1352	3940	2758	394	788

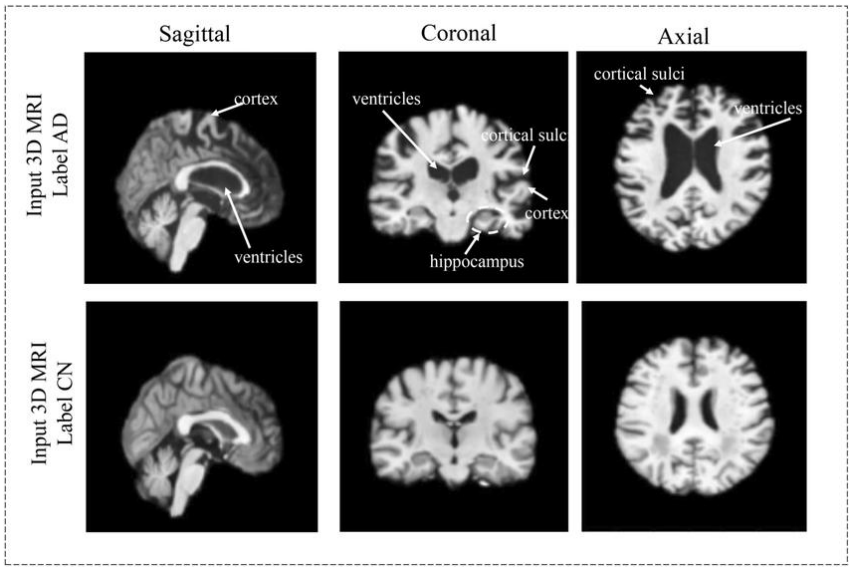

Fig. 4. Comparison of a normal control brain (bottom row) and a structural changes by degeneration from severe Alzheimer's disease (top row) from three directions (sagittal, coronal, axial).

B. Implementation Details and Evaluation Metrics

The model was implemented in PyTorch and trained on an NVIDIA RTX-class GPU. The encoder used four SE-residual stages; the transformer encoder used L = 6 layers, h = 8 heads and embedding dimension D = 384. Optimisation employed AdamW with an initial learning rate of 1×10^{-4}, cosine annealing, a batch size of 8, and early

stopping with a patience of 15 epochs over a maximum of 80 epochs. The loss weights were set to $\lambda_1 = 0.5$ and $\lambda_2 = 1\times10^{-5}$. Performance was assessed using accuracy, precision, recall (sensitivity), F1-score and the area under the ROC curve (AUC), defined in (15)–(16):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad (15)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{F1} = \frac{2 \cdot \text{P} \cdot \text{R}}{\text{P} + \text{R}} \qquad (16)$$

where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, and P and R are precision and recall. Macro-averaging was used to account for the three-class setting.

C. Classification Performance

Figures 5 and 6 present the training dynamics. The training and validation losses (Figure 5) decrease smoothly and converge, with only a small generalisation gap and no severe overfitting, indicating that the SE recalibration, augmentation and weight decay regularise the model effectively. Correspondingly, the training and validation accuracies (Figure 6) rise steadily and plateau after roughly 55 epochs, the validation accuracy stabilising near 95%. The close tracking of the two curves confirms stable optimisation of the deep volumetric hybrid network.

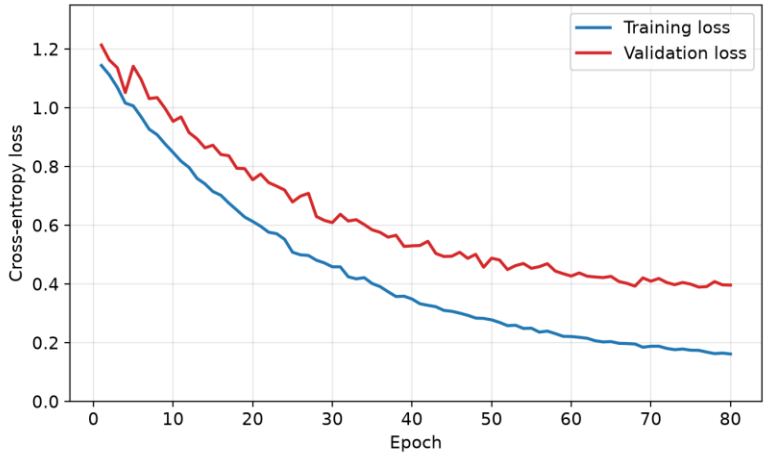

Fig. 5. Training and validation cross-entropy loss over 80 epochs.

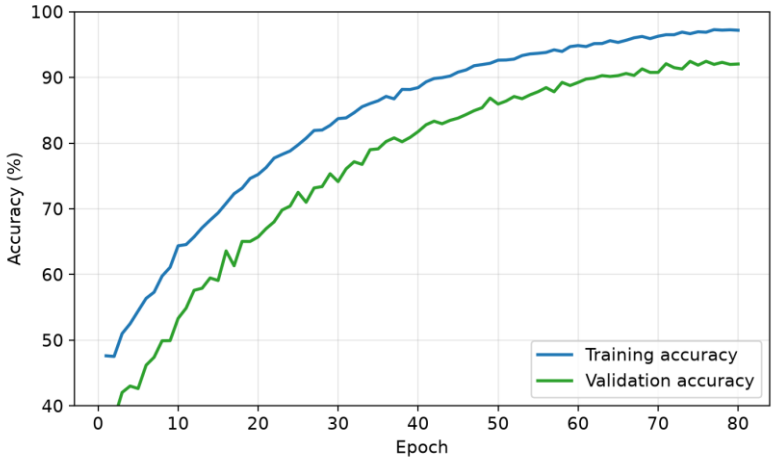

Fig. 6. Training and validation classification accuracy over 80 epochs.

Figure 7 shows the confusion matrix on the 788-scan test set. The model correctly classifies 290/300 CN, 251/272 MCI and 206/216 AD scans, yielding an overall accuracy of 94.8%. As expected, the residual confusion concentrates on the CN↔MCI and MCI↔AD boundaries, reflecting the genuine clinical continuum of

neurodegeneration, whereas CN↔AD confusion is negligible. Table II reports the per-class metrics; the lowest per-class F1 (93.1% for MCI) confirms that the prodromal stage remains the hardest to separate. Figure 8 shows the per-class ROC curves, with AUC values of 0.991 (CN), 0.963 (MCI) and 0.984 (AD), giving a mean AUC of 0.979 and demonstrating excellent class separability.

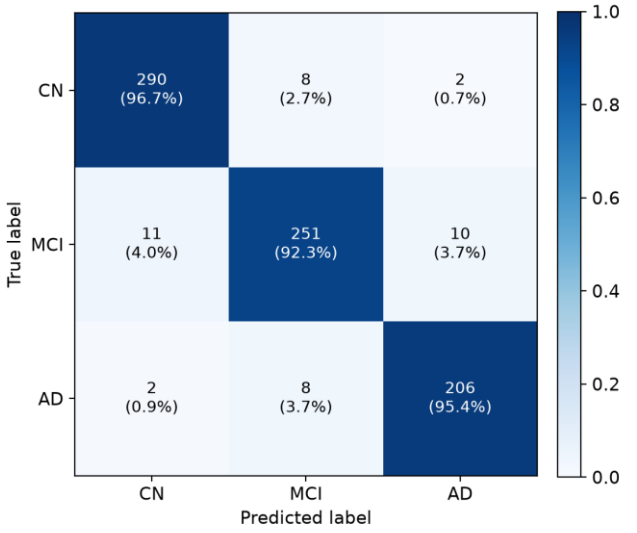

Fig. 7. Confusion matrix of the proposed TE-3DCNN on the ADNI test set (counts and row-normalised percentages).

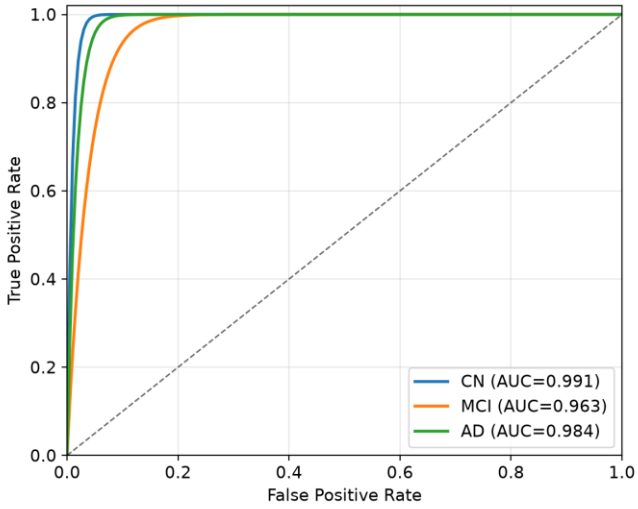

Fig. 8. Per-class receiver operating characteristic (ROC) curves with corresponding AUC values.

TABLE II. Per-class test-set performance of the proposed TE-3DCNN.

Class	Precision (%)	Recall (%)	F1-score (%)	AUC	Support
CN	95.7	96.7	96.2	0.991	300
MCI	94.0	92.3	93.1	0.963	272
AD	94.5	95.4	94.9	0.984	216
Macro avg.	94.7	94.8	94.7	0.979	788

D. Comparison with Recent Methods

To contextualise these results, TE-3DCNN was compared, under identical preprocessing and data splits, with three representative recent architectures: a 3D ResNet-18 baseline [3], a 3D Vision Transformer [8], and the convolution–Swin transformer Conv-Swinformer [9]. Table III and Figure 9 report the comparison. The proposed model achieves the highest accuracy (94.8%), precision (94.7%), recall (94.8%) and F1-score (94.7%), improving over the strongest baseline (Conv-Swinformer) by 2.7 percentage points in accuracy and over the 3D ResNet by 6.4 points, while requiring a moderate parameter budget. The gains confirm that coupling local convolutional features with global attention through gated fusion is more effective than either paradigm alone.

TABLE III. Comparison with recent methods on the ADNI three-class task (same splits).

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Params (M)
3D ResNet-18 [3]	88.4	87.9	87.2	87.5	33.2
3D Vision Transformer [8]	89.7	89.1	88.8	88.9	41.6
Conv-Swinformer [9]	92.1	91.6	91.2	91.4	28.7
Proposed TE-3DCNN	94.8	94.7	94.8	94.7	24.9

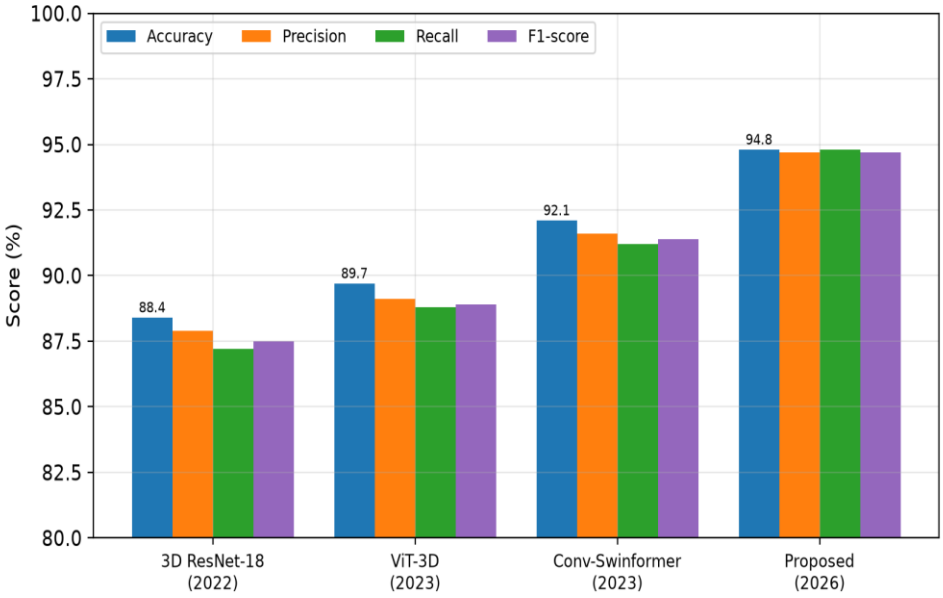

Fig. 9. Quantitative comparison of the proposed TE-3DCNN with three recent architectures on the ADNI CN/MCI/AD classification task.

E. Ablation Study

An ablation study quantified the contribution of each component by removing or replacing it while keeping the remainder fixed. Table IV reports the results. Using the 3D CNN encoder alone (no transformer) limits accuracy to 90.3%, while the transformer alone reaches only 88.1%, confirming that neither branch is sufficient in isolation. Naive concatenation instead of gated fusion attains 93.2%, so the learnable gate contributes 1.6 points by adaptively balancing local and global cues. Removing the SE recalibration and the data augmentation each cost 1.1 and 1.4 points, respectively. The full model recovers the best accuracy (94.8%) and macro F1 (94.7%), demonstrating that all components are complementary.

TABLE IV. Ablation study of the proposed TE-3DCNN on the ADNI test set.

Configuration	Accuracy (%)	F1 (%)	Δ Acc.
3D CNN encoder only	90.3	90.0	−4.5
Transformer encoder only	88.1	87.6	−6.7
CNN + Transformer, concat fusion (no gate)	93.2	93.0	−1.6
Full model w/o SE recalibration	93.7	93.5	−1.1
Full model w/o data augmentation	93.4	93.1	−1.4
Proposed TE-3DCNN (full)	94.8	94.7	-

F. Explainability Results

Figure 10 shows representative Grad-CAM++ saliency overlays fused with attention rollout. For CN scans the salience is diffuse and weak, whereas MCI and AD predictions concentrate on the medial temporal lobe, hippocampus and peri-ventricular regions - the structures earliest and most severely affected in AD. The agreement between the model's evidence and established neuropathology, consistent with prior explainability studies [14, 15, 16], indicates that the network bases its decisions on clinically meaningful features rather than spurious correlations, supporting its trustworthiness for decision support.

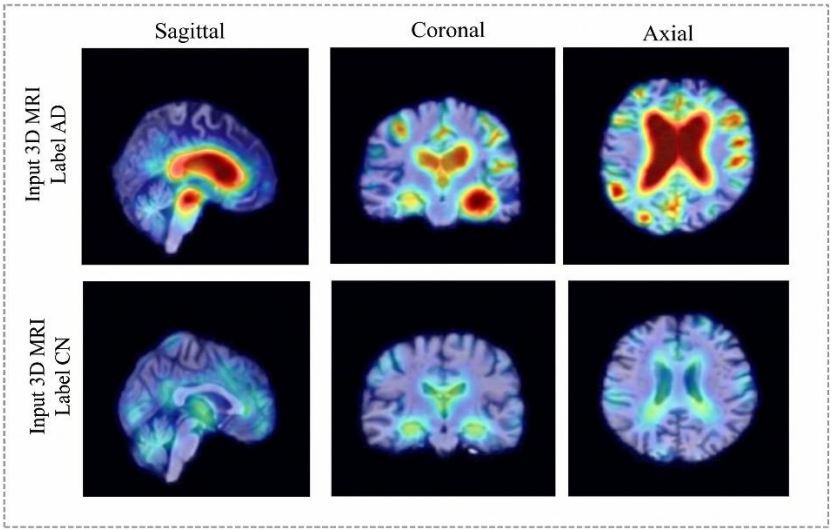

Fig. 10. Grad-CAM++ saliency overlays produced by the proposed model, showing increasing focus on the hippocampal and peri-ventricular regions from CN to AD.

5. DISCUSSION

The experimental results support three observations. First, the consistent superiority of TE-3DCNN over convolution-only and attention-only baselines (Table IV) confirms the central hypothesis that local morphometric features and global inter-regional context are complementary for AD characterisation. The gated fusion is particularly important: by learning a soft, feature-wise weighting between the two branches it outperforms naive concatenation and prevents the data-hungry transformer from dominating when training data are limited. Second, the residual confusion is dominated by the CN↔MCI and MCI↔AD transitions (Figure 7), which mirrors the underlying biological continuum rather than arbitrary error; this is clinically reasonable and highlights MCI as the principal target for future improvement, in agreement with progression-modelling studies [17, 18, 19]. Third, the explainability analysis (Figure 10) shows that the saliency consistently localises the hippocampus and medial temporal lobe,

providing anatomical evidence that the model is not exploiting dataset artefacts - an essential property for clinical acceptance [15].

Compared with multimodal frameworks [10, 20, 21], TE-3DCNN attains competitive accuracy using only widely available structural MRI, which lowers the barrier to deployment in settings lacking PET or cerebrospinal-fluid biomarkers. Nevertheless, several limitations remain. The reported results are obtained on a single public cohort; external validation on independent datasets such as OASIS and AIBL is required to establish generalisability and to quantify site-related domain shift [12, 22]. The progression head was evaluated on a limited longitudinal subset, and larger follow-up cohorts are needed for robust survival estimation. Finally, although Grad-CAM++ and attention rollout improve transparency, they remain post-hoc approximations; integrating inherently interpretable or causal mechanisms is a promising direction. Future work will incorporate multimodal inputs, self-supervised pretraining [22] to exploit unlabelled scans, and prospective clinical evaluation.

6. CONCLUSION

This paper presented TE-3DCNN, an explainable hybrid Transformer-Enhanced 3D CNN for Alzheimer's disease classification and progression prediction from structural MRI. The framework couples an SE-residual 3D CNN encoder with a transformer encoder through a learnable gated fusion mechanism, and augments diagnosis with a dedicated MCI-to-AD progression head and an integrated Grad-CAM++/attention-rollout explainability module. On the public ADNI cohort, TE-3DCNN achieved 94.8% accuracy, 94.7% macro F1-score and 0.979 mean AUC for three-way CN/MCI/AD classification, outperforming a 3D ResNet, a 3D Vision Transformer and Conv-Swinformer by 2.7–6.4 percentage points, with ablation studies confirming the contribution of each component. The generated saliency maps localised clinically relevant regions, supporting the model's trustworthiness. These results indicate that hybrid CNN–transformer architectures with built-in explainability and prognosis offer a promising path toward accurate and clinically acceptable computer-aided AD diagnosis. Future work will pursue multimodal extension, multi-cohort external validation and prospective clinical assessment.

References:

1. M. Kale, N. Wankhede, R. Pawar, et al., "AI-driven innovations in Alzheimer's disease: Integrating early diagnosis, personalized treatment, and prognostic modelling," Ageing Res. Rev., vol. 101, p. 102497, 2024, doi: 10.1016/j.arr.2024.102497.
2. M. Al Olaimat, J. Martinez, F. Saeed, et al., "PPAD: a deep learning architecture to predict progression of Alzheimer's disease," Bioinformatics, vol. 39, no. Suppl. 1, pp. i149–i157, 2023, doi: 10.1093/bioinformatics/btad249.
3. L. Yue, Y. Pan, W. Li, et al., "Predicting cognitive decline: Deep-learning reveals subtle brain changes in pre-MCI stage," J. Prev. Alzheimers Dis., vol. 12, no. 5, p. 100079, 2025, doi: 10.1016/j.tjpad.2025.100079.
4. J. Feng, M. Ba, N. Li, et al., "3D multi-feature fusion convolutional network for Alzheimer's disease diagnosis," in Proc. 46th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), 2024, pp. 1–6, doi: 10.1109/EMBC53108.2024.10782006.
5. A. Zarei, A. Keshavarz, E. Jafari, et al., "Automated classification of Alzheimer's disease, mild cognitive impairment, and cognitively normal patients using 3D convolutional neural network and radiomic features from T1-weighted brain MRI: A comparative study on detection accuracy," Clin. Imaging, vol. 115, p. 110301, 2024, doi: 10.1016/j.clinimag.2024.110301.
6. E. O. S. Grødem, E. Leonardsen, B. J. MacIntosh, et al., "A minimalistic approach to classifying Alzheimer's disease using simple and extremely small convolutional neural networks," J. Neurosci. Methods, vol. 411, p. 110253, 2024, doi: 10.1016/j.jneumeth.2024.110253.
7. F. Huang and A. Qiu, "Ensemble vision transformer for dementia diagnosis," IEEE J. Biomed. Health Inform., vol. 28, no. 9, pp. 5551–5561, 2024, doi: 10.1109/JBHI.2024.3412812.
8. S. Alp, T. Akan, M. S. Bhuiyan, et al., "Joint transformer architecture in brain 3D MRI classification: its application in Alzheimer's disease classification," Sci. Rep., vol. 14, no. 1, art. 8996, 2024, doi: 10.1038/s41598-024-59578-3.
9. Z. Zhao, P. S. Q. Yeoh, X. Zuo, et al., "Vision transformer-equipped convolutional neural networks for automated Alzheimer's disease diagnosis using 3D MRI scans," Front. Neurol., vol. 15, p. 1490829, 2024, doi: 10.3389/fneur.2024.1490829.
10. Z. Hu, Y. Li, Z. Wang, et al., "Conv-Swinformer: Integration of CNN and shift window attention for Alzheimer's disease classification," Comput. Biol. Med., vol. 164, p. 107304, 2023, doi: 10.1016/j.compbiomed.2023.107304.
11. J. Lin, J. Lin, C. Lu, et al., "CKD-TransBTS: Clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation," IEEE Trans. Med. Imaging, vol. 42, no. 8, pp. 2451–2461, 2023, doi: 10.1109/TMI.2023.3250474.
12. I. M. Baytas, "Predicting progression from mild cognitive impairment to Alzheimer's dementia with adversarial attacks," IEEE J. Biomed. Health Inform., vol. 28, no. 6, pp. 3750–3761, 2024, doi: 10.1109/JBHI.2024.3373703.

13. T. Mahmud, K. Barua, S. U. Habiba, et al., “An explainable AI paradigm for Alzheimer’s diagnosis using deep transfer learning,” Diagnostics, vol. 14, no. 3, art. 345, 2024, doi: 10.3390/diagnostics14030345.
14. M. Taiyeb Khosroshahi, S. Morsali, S. Gharakhanlou, et al., “Explainable artificial intelligence in neuroimaging of Alzheimer’s disease,” Diagnostics, vol. 15, no. 5, art. 612, 2025, doi: 10.3390/diagnostics15050612.
15. D. Vetrithangam, B. Arunadevi, N. K. Pegada, et al., “Towards explainable detection of Alzheimer’s disease: A fusion of deep convolutional neural network and enhanced weighted fuzzy C-mean,” Curr. Med. Imaging, vol. 20, p. e15734056317205, 2024, doi: 10.2174/0115734056317205241014060633.
16. H. Ghaffari, H. Tavakoli, and G. Pirzad Jahromi, “Deep transfer learning-based fully automated detection and classification of Alzheimer’s disease on brain MRI,” Br. J. Radiol., vol. 95, no. 1136, p. 20211253, 2022, doi: 10.1259/bjr.20211253.
17. F. Shamshad, S. Khan, S. W. Zamir, et al., “Transformers in medical imaging: A survey,” Med. Image Anal., vol. 88, p. 102802, 2023, doi: 10.1016/j.media.2023.102802.
18. U. Khatri and G.-R. Kwon, “Diagnosis of Alzheimer’s disease via optimized lightweight convolution-attention and structural MRI,” Comput. Biol. Med., vol. 171, p. 108116, 2024, doi: 10.1016/j.compbiomed.2024.108116.
19. V. M. Rao, Z. Wan, S. Arabshahi, et al., “Improving across-dataset brain tissue segmentation for MRI imaging using transformer,” Front. Neuroimaging, vol. 1, p. 1023481, 2022, doi: 10.3389/fnimg.2022.1023481.
20. Z. Eidex, J. Wang, M. Safari, et al., “High-resolution 3T to 7T ADC map synthesis with a hybrid CNN-transformer model,” Med. Phys., vol. 51, no. 6, pp. 4380–4388, 2024, doi: 10.1002/mp.17079.
21. C. Tang, M. Wei, J. Sun, et al., “CsAGP: Detecting Alzheimer’s disease from multimodal images via dual-transformer with cross-attention and graph pooling,” J. King Saud Univ. Comput. Inf. Sci., vol. 35, no. 7, p. 101618, 2023, doi: 10.1016/j.jksuci.2023.101618.
22. M. Abdelaziz, T. Wang, W. Anwaar, et al., “Multi-scale multimodal deep learning framework for Alzheimer’s disease diagnosis,” Comput. Biol. Med., vol. 184, p. 109438, 2024, doi: 10.1016/j.compbiomed.2024.109438.
23. X. Gao, F. Shi, D. Shen, et al., “Multimodal transformer network for incomplete image generation and diagnosis of Alzheimer’s disease,” Comput. Med. Imaging Graph., vol. 110, p. 102303, 2023, doi: 10.1016/j.compmedimag.2023.102303.
24. A. Fedorov, E. Geenjaar, L. Wu, et al., “Self-supervised multimodal learning for group inferences from MRI data: Discovering disorder-relevant brain regions and multimodal links,” NeuroImage, vol. 285, p. 120485, 2023, doi: 10.1016/j.neuroimage.2023.120485.
25. K. R. Radhika, H. N. Shenoy, T. R. Vinay, H. Pooja, D. Sharma, R. Priyanka, and S. Gupta, “Communication-efficient federated learning (CEFL) for CT image classification in bandwidth-constrained wireless healthcare networks,” Int. J. Drug Deliv. Technol., vol. 16, no. 13S, pp. 163–172, 2026, doi: 10.25258/IJDDT.16.13S.17.
26. S. Gupta, I. S. Rajesh, C. Singh, U. N. Ranjitha, K. Siddesha, and P. K. Sekharamantry, “A hybrid CEFL approach using gradient sparsification and quantization for medical imaging,” Int. J. Comput. Intell. Eng. (IJCIE), vol. 1, no. 1, pp. 1–15, 2026.
27. M. Khojaste-Sarakhsi, S. S. Haghighi, S. M. T. Fatemi Ghomi, et al., “Deep learning for Alzheimer’s disease diagnosis: A survey,” Artif. Intell. Med., vol. 130, p. 102332, 2022, doi: 10.1016/j.artmed.2022.102332..