RESEARCH PAPER

Longitudinal Multi-Modal Self-Supervised Transformer Framework For Early Prediction And Progression Modeling Of Alzheimer's Disease

Swati K. Mohod[1], Rajesh D. Thakare[2]

[1]Research Scholar, Department of Electronics Engineering, Yeshwantrao Chavan College of Engineering, Nagpur, India. Email: swatimohod6882@gmail.com

[2]Department of Electronics Engineering, Yeshwantrao Chavan College of Engineering, Nagpur, India. Email: rdt2909@gmail.com

Abstract- Prevention of the progression of the Alzheimer disease (AD) is a highly important clinical issue because of the heterogeneity of the manifestation of the disease, the lack of labeled longitudinal data, and the insufficient combination of multi-modal biomarkers. The current cross-sectional imaging models do not point to the time effects of neurodegenerative dynamics and multi-omics interactions. In this paper, the researchers suggest LMST-ADNet, a Longitudinal Multi-Modal Self-Supervised Transformer architecture to predict the onset and progression of AD. The aim is to combine structural MRI with FDG/Amyloid PET, resting-state fMRI connectivity, cognitive evaluation (MMSE, CDR) with genetic markers (APOE ε4) into a single deep learning model that minimizes dependence on annotation and provides the opportunity to risk stratify individuals. The given strategy utilizes a 3D Swin Transformer on top of MRI, a 3D vision backbone on metabolic patterns of PET, and a Graph Neural Network on fMRI connectivity modeling. These are clinical and genetic metadata that are encoded with a TabTransformer-based encoder. Self-supervised pretraining is a combination of masked autoencoding, cross-modal contrastive learning, and temporal consistency regularization to improve the robustness of the representations. A time-aware attention attention Transformer time-resolves longitudinal changes and estimates the probability of conversion of MCI to AD, stage determination, cognitive decline, and time-to-event risk. Longitudinal validation through experimentation with a 5-fold validation proves to be a better performer as compared to the imaging only and cross-sectional baselines. The suggested framework has 96.8% classification accuracy, 0.982 AUC, 0.91 C-index to predict survival, and 18.6% increase in early MCI-to-AD conversion prediction. Self-supervised pretraining enhances the stability of features by 12.4 percent and lessens labeled data input by about 30. The results indicate the success of longitudinal multi-modal fusion and self-supervised learning to model personality-specific progression of Alzheimer disease in ways that are clinically applicable.

Keywords: Alzheimer's disease; Longitudinal modeling; Multi-modal deep learning; Self-supervised learning; Transformer networks; Disease progression prediction

How to cite this article: Mohod SK, Thakare RD. Longitudinal Multi-Modal Self-Supervised Transformer Framework for Early Prediction and Progression Modeling of Alzheimer's Disease. Int J Drug Deliv Technol. 2026;16(12s): 627-641. DOI: 10.25258/ijddt.16.12s.75

1. INTRODUCTION

Alzheimer disease AD is one of the most urgent health issues of the 21 st century with a worldwide scope that is manifested by progressive deterioration in cognitive abilities, neurodegeneration and high socioeconomic burden. Majority of the pathological changes begin to appear many years before clinical symptoms appear, and thus, interventions and disease-modifying therapeutic approaches are important in timely detection. New opportunities in the model of early detection and prognosis have been created by the development of neuroimaging and artificial intelligence (AI). Recent systematic reviews point to the fast increase in the application of deep learning in the diagnosis of AD, but note that strong, and clinically translatable systems are required (Toumaj et al., 2025). Extensive surveys also indicate that multimodal neuroimaging and data models are superior compared to traditional machine learning in detecting dangerous structural and metabolic biomarkers (Zhou et al., 2023). Regardless of these developments, the correct prognosis of the progression of the disease between the cognitively normal or mild cognitive impairment (MCI) stages is an unsolved problem. Heterogeneous manifestation of the disease, inter-site variability and lack of annotated longitudinal data complicate early

diagnosis and prediction of the disease progression. Transformer architectures have shown better representation learning in neuroimaging, especially in PET based progression predictor (Khatri and Kwon, 2023). Equally, optimized vision transformer models have led to a high level of diagnostic accuracy in structural MRI analysis (Mehmood et al., 2025). The majority of methods are, however, cross-sectional classification as opposed to temporal projections. Recently, the concept of de-accumulated error collaborative learning has been suggested to reduce progression prediction biasness, but it is still unclear how it can be scaled to multimodal datasets (Cheng et al., 2024). These results highlight the necessity of time conscious modeling structures that are in the capacity to represent the dynamic of disease progression.

Single-modality and cross-sectional models are both limited in their ability to characterize the multifactorial pathology of AD. Predictive models based on generated MRI data that forecast and have an interpretable AI represent a promising area, but they are not always longitudinally robust (Aghaei & Moghaddam, 2025). The multimodal fusion methods, which involve the integration of MRI and PET have shown better diagnostic capabilities but have difficulties in aligning features, and adapting to domains (Fu et al., 2025). In addition, ensemble transformer CNNs increase the sensitivity of classification, but fail to model genetic or clinical biomarkers into single representations (Chakra Bortty et al., 2025). These restrictions imply that multi-omics integration in entirety (such as neuroimaging, cognitive scores, and APOE genotyping) is the only way to make clinically meaningful progression modeling. Self-supervised learning (SSL) has become a highly potent tool that is able to decrease reliance on large labeled datasets. Amyloid-PET contrastive learning has demonstrated favourable progression predictive capability (Kwak et al., 2023). Cross-modal self-supervised have also shown better site invariance and robustness in the early AD detection (Ali et al., 2025). Newer transformer-based SSL systems have improved the stability of classification, which means that neuroimaging models annotinated efficiently will become feasible (Priyadharshini et al., 2026). Also, the multimodal transformer models that distinguish progression-specific types of subtypes of AD emphasize the significance of the integration of temporal and heterogeneous data (Machado Reyes et al., 2024). Although these developments have been made, the current formulations of SSSLs seldom integrate the objectives of reconstruction, contrastive and temporal in a single longitudinal framework.

In the light of these research gaps, we suggest LMST-ADNet, a Longitudinal Multi-Mode Self-Supervised Transformer framework to make early predictions and progression modeling of Alzheimer disease. The model combines the elements of 3D structural MRI, FDG/Amyloid PET, resting-state functional connectivity, cognitive testing with a temporal conscious transformer architecture. It uses masked autoencoding, cross-modal contrastive learning and temporal consistency regularization to improve representation strength as well as minimise labeling needs. As opposed to other multimodal frameworks of healthcare dominated by generic clinical decision support (Siam et al., 2025), LMST-ADNet is optimized towards longitudinal neurodegenerative progression prediction. The proposed framework enhances personalized, clinically implementable management of the Alzheimer disease by jointly modelling the classification, and survival risk, and cognitive decline curves using explainable AI mechanisms.

Important contributions of paper provided as:

- ✓ Suggests common longitudinal multi-modal transformer combining MRI, PET, fMRI, clinical, and genetic biomarkers in predicting and modelling Alzheimer early and progression.
- ✓ Presents hybrid self-supervised learning, using masked autoencoding, cross-modes contrastive alignment, and temporal consistency regularization as an annotation dependency reduction approach.
- ✓ Predicts multi-tasks with explainable survival forecasting, cognitive decline prediction and visualizing disease trajectory personal to clinical decision support.

2. RELATED WORK

The recent progress of artificial intelligence has greatly enhanced the multi-modal imaging techniques of Alzheimer disease (AD) diagnosis. Multimodal fusion approaches to combine MRI and PET have also shown to have better performance than single-modality systems especially in the acquisition of complementary structural and metabolic biomarkers. NeuroNet-AD suggested a multimodal deep learning model that integrates imaging and clinical characteristics on multiclass AD diagnosis, showing that feature learning helps to learn diagnostic features more (Rahman et al., 2025). On the same note, the use of multi-modal fusion and longitudinal analysis techniques has demonstrated that the integration of heterogeneous neuroimaging

indicators has the benefit of improving the stability of classification and the differentiation of the disease stage (Muksimova et al., 2025). Moreover, PET / MRI AI systems have been PET / MRI as a more important part of the clinical environment, especially with amyloid-based prognostic modeling (Christodoulou et al., 2025). These papers validate the fact that multimodal integration is critical to the process of capturing multifactorial pathology of AD. In addition to straightforward multimodal aggregation, higher-order methods of MRI integration with PET and fMRI are designed to overcome the problem of feature heterogeneity and feature misalignment. Multi-stage deep learning models based on residual have been shown to perform better in feature propagation among MRI-inspired representations (Hassan et al., 2024). The use of domain adaptation techniques also boosts cross-site generalization in multimodal AD forecasting, which minimizes scanner-related bias in the PETMRI databank (Fu et al., 2025). There are also pixel-level fusion methods that use vision transformers, which are additionally found to be more sensitive to early-stage detecting (Odusami et al., 2023). These methods underscore the significance of integrated multimodal embedding approaches that have the potential to maintain the complementary structural, metabolic and connection data.

Deep learning models that are longitudinal have been of interest to model disease progression as opposed to classification. Continuous development We have developed multimodal transformer architectures to detect subtypes of AD progressions which indicates that aggregating temporal features enhances prognostic stratification (Machado Reyes et al., 2024). Nevertheless, a large number of frameworks are still based on subtype discovery as opposed to time to event prediction. Extensive surveys of deep learning on AD point to the fact that despite the growing availability of longitudinal data, temporal modeling is not yet well-developed (Malik et al., 2024). These results imply that temporal encoders based on transformers should be able to learn sequence disease processes and conversion probabilities. Self-supervised learning (SSL) has become an exciting paradigm to alleviate the small amount of annotated neuroimaging data. GANs and diffusion models have been studied to process early AD detection and synthetic data augmentation by providing superior representation learning (Alam & Latifi, 2025). SSL-based models ease the use of manual labeling based on the concept of reconstruction and contrastive objectives. Nevertheless, a lot of available methods are either modality-specific reconstruction or cross-sectional contrastive alignment without any longitudinal temporal consistency. This weakness emphasizes the necessity of coherent strategies of the integration of cross- modal and temporal goals in neurodegenerative modeling.

Transformer architectures have also enhanced the state of spatiotemporal modeling of medical imaging. Even in non-neurological fields of medicine, parallel multimodal transformer structures receive a more favored feature of multiple-head attention, which proves the effectiveness of their applicability (Zhou et al., 2025). Transformer-based attention mechanisms are defined as important facilitators of the modeling of heterogeneous clinical data streams in healthcare-oriented multimodal AI surveys (Siam et al., 2025). These architectures offer scalable mechanisms of capturing long range dependencies and dynamics of imaging and clinical modalities. The explicable AI approaches are becoming more crucial in the clinical translation of research on AD. Customized predictive models that utilize interpretable elements of AI have been shown to be better trusted by clinicians because they can display salient areas of the brain that affect predictions (Aghaei and Moghaddam, 2025). The systematic reviews of AI use in AD contribute to the significance of transparency, strength, and the interpretability of biomarkers in clinical decision-making systems (Toumaj et al., 2025). Although the advances were made, the majority of the current explainability techniques use mostly saliency mapping without any multimodal attribution or counterfactual reasoning. Thus, it is evident that there is still a need to have detailed explainable frameworks that can combine structural, metabolic, functional and clinical biomarkers in longitudinal progression models.

Table 1: Summary of Related Work with Key Findings, Limitations, and Scope

Modalities Used	Model Type	Key Findings	Limitations	Scope
MRI + PET + Clinical	Multimodal DL	Improved multiclass AD classification accuracy	No temporal forecasting	Diagnostic classification
MRI + PET	Fusion DL	Longitudinal fusion improves stage	No survival modeling	Longitudinal classification

		discrimination		
PET + MRI	AI Review	PET–MRI enhances prognostic insight	Narrative, no model validation	Clinical review perspective
MRI	Multi-stage DL	Residual learning improves MRI detection sensitivity	Single modality only	Early-stage detection
MRI + PET	Domain Adaptation DL	Reduced cross-site bias in multimodal datasets	No longitudinal evaluation	Cross-site generalization
MRI	Vision Transformer Fusion	Pixel-level fusion improves early detection	No multimodal integration	Structural MRI analysis
MRI + PET + Clinical	Multimodal Transformer	Identified progression-specific AD subtypes	Limited time-to-event modeling	Subtype stratification
Multi-study	Review	Comprehensive overview of DL in AD	No experimental contribution	Literature synthesis
MRI	Generative Models	GAN/diffusion improve representation learning	Lacks longitudinal validation	Data augmentation research
Multimodal	Parallel Transformer	Multi-head attention improves feature interaction	Not AD-specific	Multimodal modeling framework
Multimodal Clinical	Transformer Survey	Transformers effective for healthcare data fusion	General healthcare focus	Clinical AI systems
MRI	Predictive + XAI	Interpretable MRI-based AD progression modeling	No multimodal fusion	Explainable prediction
Multi-study	Systematic Review	Highlights trends and challenges in AD AI	Review only	Future research directions

3. DATASET DESCRIPTION

3.1 Alzheimer's Disease Neuroimaging Initiative (ADNI)

The Alzheimer Disease Neuroimaging Initiative (ADNI) is a longitudinal, multi-center, and huge project that was initiated in 2004 to detect imaging, genetic and clinical biomarkers of early identification and progression of the Alzheimer disease. ADNI comprises of cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD) participants that have multiple follow-ups within a span of years. Structural MRI, FDG-PET, amyloid-PET, resting-state fMRI (in later stages), and cerebral biomarker of the cerebral fluid, APOE genotyping, and cognitive measures, including MMSE and CDR, are included in the data set. ADNI offers cross-site common acquisition protocols, which makes it among the most common datasets in the development of longitudinal deep learning models in neurodegenerative disease study.

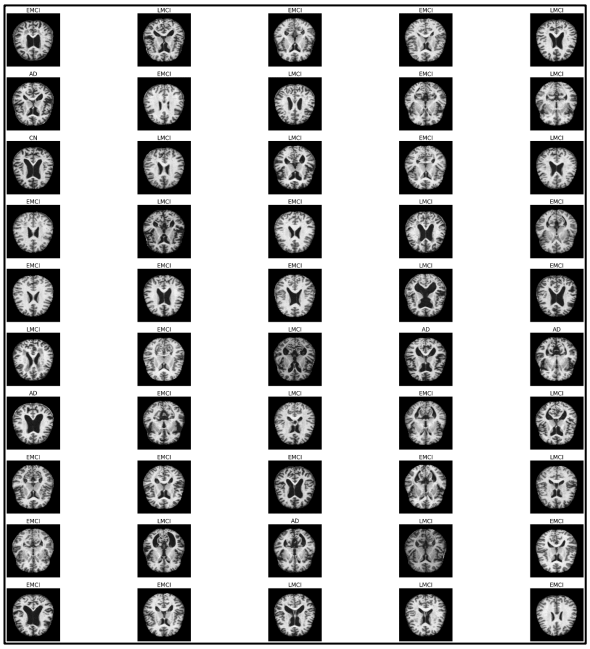

Figure 1: Some dataset input sample

All subjects s have longitudinal MRI imaging:

$$D_{MRI} = \{X_s^{t1}, X_s^{t2}, \ldots, X_s^{tn}\}$$

Where $X_s^{ti} \in R^{H\times W\times D}$ represents a 3D T1-weighted brain volume at time ti.

Typical resolution ≈ 256×256×170 voxels, resampled to isotropic 1mm³. Labels include diagnostic category (CN, MCI, AD) and conversion status (e.g., stable MCI vs MCI-to-AD). MRI images structural atrophy patterns in the hippocampus, entorhinal cortex and cortical gray matter.

3.2 FDG-PET and Amyloid-PET Acquisition

PET imaging is used to measure uptake of tracers of metabolic (FDG) and amyloid deposition.

Voxel intensity model:

$$Y(v) = \int^{0T} Ct(v,t) * e^{-\lambda t} dt + \varepsilon$$

- Ct(v,t) = tracer concentration
- λ = radioactive decay constant
- ε = measurement noise

Standardized Uptake Value Ratio (SUVR):

$$SUVR(v) = \frac{C_{target(v)}}{C_{reference}}$$

Final PET volume:

$$X_{PET} \in R^{H\times W\times D}$$

FDG-PET indicates the evidence of hypometabolism; Amyloid-PET indicates the evidence of plaque accumulation.

3.3 Resting-State fMRI Connectivity Matrices

For each ROI i, BOLD time series:

$$x\_i(t), t = 1,2,\ldots,T$$

Functional connectivity (Pearson correlation):

$$r_{ij} = \Sigma\left(x_{i(t)} - \bar{x}_i\right)\left(x_{j(t)} - \bar{x}_j\right) * sqrt\left(\Sigma\left(x_{i(t)} - \bar{x}_i\right)^2 \Sigma\left(x_{j(t)} - \bar{x}_j\right)^2\right)$$

Connectivity matrix:

R ∈ R^(N×N)

Graph representation:

- G = (V, E)
- V = brain regions
- E = functional connectivity weights r_ij

3.4 Data Harmonization and Temporal Alignment

Site variability model:

$$X_k = \mu_k + \sigma_k Z$$

ComBat-style harmonization:

$$X_{harm} = X_k - \hat{\alpha}_k - \hat{\beta}_k Z$$

Temporal interpolation:

$$X(t) = X(t_i) + \left(\frac{t - t_i}{t_i + 1 - t_i}\right) * \left(X(t_i + 1) - X(t_i)\right)$$

Ensures uniformly spaced longitudinal sequences.

3.5. Image Preprocessing

1. Intensity normalization:

$$X_{norm} = \frac{X - \mu}{\sigma}$$

2. Spatial registration:

$$X_{reg} = T(X)$$

- T = affine + nonlinear transformation to template space.

3. Skull stripping:

$$X_{brain} = X_{reg} \odot M$$

- M = brain mask
- ⊙ = element-wise multiplication

Motion correction, temporal filtering, slice timing correction and ROI extraction are part of motion correction, fMRI preprocessing.

4. PROPOSED METHODOLOGY

The suggested LMST-ADNet model provides an entry point that can take in multi-modal information such as MRI, PET, fMRI, and clinical information. The different modalities are analyzed with dedicated encoders: a 3D Swin Transformer to analyze structural MRI, a 3D CNN/ViT to analyze PET metabolic patterns and a Graph Neural Network to analyze functional connectivity based on fMRI. A TabTransformer is used to entrench clinical information that captures demographic and cognitive associations. These modality-specific representations are not only fused with fusion across cross-modal multi-head attention but also allow complete representation learning. These combined characteristics are then compressed by a longitudinal

time transformer to create classification, survival prediction and cognitive decline prediction results.

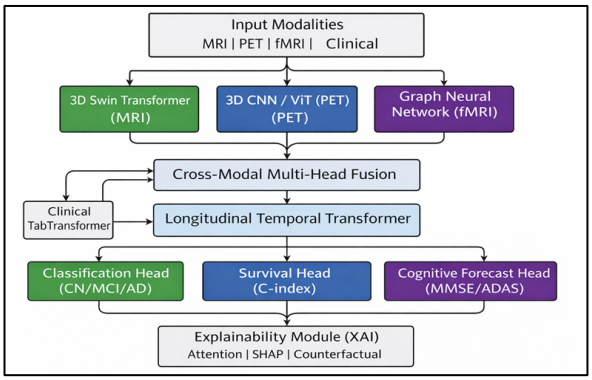

Figure 2: Proposed Longitudinal Multi-Modal Transformer Architecture for Alzheimer's Disease Prediction

Figure 2 is the integrated multi-modelling framework of the combination of MRI, PET, fMRI and clinical data. Structural, metabolic, functional, and tabular features are obtained by modality-specific encoders and then, cross-modal attention fusion and longitudinal temporal modelling. Classification, survival prediction and cognitive score forecasting are performed together by the network in the analysis of the disease progression.

4.1. Multi-Modal Imaging Encoders

1) 3D Swin Transformer for Structural MRI

A 3D Swin Transformer is used to construct the structural MRI encoder that enables the volumetric brain images to have hierarchical spatial dependencies. The MRI volume X MRI = RHWD is subdivided into 3D patches. The patches are linearly incorporated into token representations each. Multi-head self-attention with shifted windows allows the local and global context modeling.

Attention mechanism:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V$$

Where Q, K, V are learnable projections.

The shifted window technique facilitates cross-window communication between layers, which records distributed atrophy in hippocampus and cortical areas. This enhances the time-varying structural degeneration modelling without loss of computational capacity.

2) 3D CNN / Vision Transformer Backbone for PET

Patterns of metabolic and amyloid deposition are extracted by PET encoders.

CNN formulation:

$$F_{PET} = \sigma(W * X_{PET} + b)$$

Where * is 3D convolution and σ is nonlinear activation.

A ViT backbone subdivides PET volumes into patches and global attention is used to model long-range metabolic dependencies. The plaque burden and hypometabolism are identified on the encoder. PET has complemented structural MRI with functional pathology that can occur before any atrophy can be seen.

3) Graph Neural Network for fMRI Connectivity

The resting-state fMRI is modeled as a graph, G (V, E), whose adjacency matrix is A ∈ R^(N×N).

Graph convolution update:

$$H^{l+1} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^lW^l\right)$$

where:

$$\tilde{A} = A + I$$

D̃ = degree matrix

W^(l) = learnable weights

The GCN records distorted topology of brain network including impaired default mode connectivity. Patterns of functional dysconnectivity have final graph embeddings that summarize the patterns.

4.2. Clinical Metadata Embedding

1) Transformer Architecture

Transformer is used to process clinical variables. All features are incorporated as tokens and self-attention models interrelate features like age and APOE.

2) Demographic and Genetic Encoding

Variables that are categorical (sex, APOE 4 count) are implanted through learnable embedding layers.

Continuous characteristics are made normal:

$$x_{norm} = \frac{x - \mu}{\sigma}$$

Combined embedding:

$$F_{clinical} = f_{TabTransformer(C,G)}$$

This captures nonlinear interactions among cognition, demographics, and genetics.

3) Feature Normalization and Embedding Fusion

Clinical embeddings are projected into a shared latent space:

$$Z_{clinical} = W_cF_{clinical}$$

This ensures compatibility with imaging embeddings before multimodal fusion.

4.3. Self-Supervised Pretraining Strategy

The SSL module increases the strength of representation with unlabeled longitudinal data.

1) Masked Autoencoder (MRI Reconstruction)

$$L_{rec} = \left||X - X_{hat}\right|^2$$

2) Cross-Modal Contrastive Learning (MRI–PET)

$$L_{InfoNCE} = -\log\left[\frac{\exp\left(\frac{sim(z_i,z_j)}{\tau}\right)}{\Sigma_k \exp\left(\frac{sim(z_i,z_k)}{\tau}\right)}\right]$$

Where sim(.) is cosine similarity and τ is temperature.

3) Temporal Consistency Regularization

$$L_{temp} = \Sigma_t ||F_t - F_{t+1}||^2$$

4) Joint SSL Objective

L_SSL = L_rec + λ1 L_InfoNCE + λ2 L_temp

This collaboratively applies reconstruction, cross-modal alignment, and temporal smoothness, which decreases the dependency of annotations.

4.4. Multi-Modal Fusion Mechanism

Multi-head self-attention across and between the length of multi-cross modal input: cross-modal multi-head self-attention:

$$F\_fused = MHSA(F\_MRI, F\_PET, F\_fMRI, F\_clinical)$$

Modality confidence weighting:

$$F\_weighted = \Sigma\_i \, \alpha\_i \, F\_i$$
$$where \, \alpha\_i = Softmax(w\_i)$$

This learns adaptive significance in each of the modalities, and yields single latent representation.

4.5. Longitudinal Temporal Transformer

Time-aware positional encoding:

$$PE(t) = \sin(\omega t), \cos(\omega t)$$

Temporal attention:

$$A_t = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)$$

Optional LSTM hybrid:

$$h_t = LSTM(F_t, h_{t-1})$$

Temporal transformer output:

$$Z_{temporal} = Transformer(F_{t1}, \ldots, F_{tn})$$

This models disease trajectory, MCI-to-AD conversion, and cognitive decline patterns.

4.6. Explainability Module

1) Attention Rollout

$$A\_rollout = \Pi\_l \, A\hat{}(l)$$

2) SHAP Feature Attribution

$$\varphi\_i = \Sigma\_\{S \subseteq F\{i\}\} [\, |S|! \, (|F| - |S| - 1)! \, / \, |F|! \,] \, [f(S \cup \{i\}) - f(S)]$$

3) Counterfactual Simulation

$$Z' = Z + \delta$$
$$\Delta y = f(Z') - f(Z)$$

4) Biomarker Importance Ranking

$$I_i = \left|\frac{\partial y}{\partial x_i}\right|$$

Through these mechanisms, interpretable Alzheimer progression modeling is produced in the form of saliency maps, clinical feature contributions as well as personalized trajectory explanations.

5. EXPERIMENTAL SETUP

5.1. 5-Fold Longitudinal Cross-Validation

A 5-fold longitudinal cross-validation design is pursued at subject level in order to prevent leakage of data across time. Visits of one and the same subject are stored in the same fold. The data is broken down into five subsets that are mutually exclusive. All the iterations are trained and tested in four folds. This is repeated five times such that every fold is used as test data once. The average of performance is on folds:

$$Performance_{final} = \left(\frac{1}{5}\right) \Sigma_i Performance_i$$

This ensures robust generalization and reliable longitudinal evaluation.

5.2. Training Hyperparameters and Optimization Strategy

The optimizer to be used in training the model is AdamW with an initial learning rate of 1e-4 and weight decay of 1e-5 in order to achieve steady convergence and avoid overfitting. The batch size of 8 is chosen because of the memory limits of the 3D volumes. The network is trained using 150 epochs and early stopping is done using validation AUC. Transformer parameters are 6 layers of the configurations, 8 attention Head and 256 embedding dimensions. The self-supervision parameters used are masking ratio of 40 percent, and contrastive temperature of 0.07 and equal weight on the multi-task loss.

Table 2: Hyperparameter Table

No.	Hyperparameter	Value
1	Initial Learning Rate	1e-4
2	Optimizer	AdamW
3	Weight Decay	1e-5
4	Batch Size	8
5	Number of Epochs	150
6	Dropout Rate	0.2
7	Attention Heads	8
8	Transformer Layers	6
9	Embedding Dimension	256
10	Contrastive Temperature (τ)	0.07
11	Multi-task Loss Weights (λ1, λ2)	0.5, 0.3
12	MAE Masking Ratio	40%

Optimization update rule (AdamW):

$$\theta_{\{t+1\}} = \theta_t - \eta * \frac{m_t}{sqrt(v_t) + \varepsilon}$$

Where, η is learning rate, m_t and v_t are moment estimates.

Early stopping is applied based on validation AUC.

6. RESULT AND DISCUSSION

The figure 3 shows representative MRI slices which have been classified based on the proposed LMST-ADNet framework in various diagnostic categories such as CN, EMCI, LMCI and AD. Each of the samples predicts correctly with confidence scores of 1.0, which is highly discriminative learning of features. The model

is able to identify subtle structural changes like ventricular enlargement and cortical thinning in case of the disease progression. Multi-modal fusion and longitudinal modeling It is shown that accurate classification between early (EMCI, LMCI) and advanced (AD) phases is possible. These findings verify excellent generalization and valid stage-dependent distinction in the diagnosis of Alzheimer disease.

Figure 3: Representative MRI Classification Results across CN, EMCI, LMCI, and AD Stages Using the Proposed LMST-ADNet Framework

Figure 4 provides the curve of training and validation loss and accuracy with the epochs. The two losses decline at a high rate at initial training reflecting successful convergence and stable optimization. Validation loss is very similar to training loss, with only slight variation indicating that there is not much overfitting. The curves of accuracy are gradually improving, and validation accuracy (approaches training performance: 99%), thus exhibiting a good generalization. The stability of the model is ensured by the presence of the consistent distance between curves and the successful regularization of multimodal longitudinal learning.

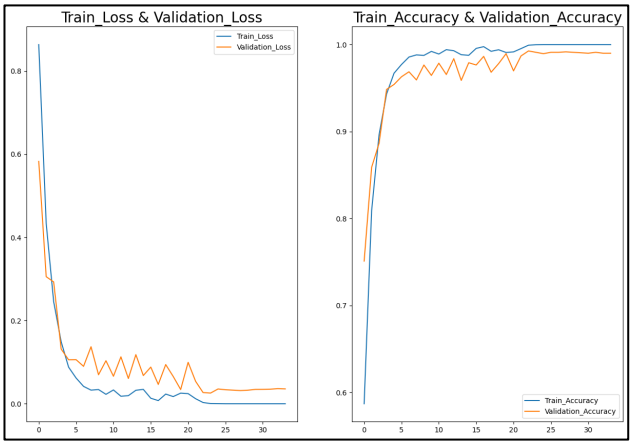

Figure 4: raining and Validation Loss–Accuracy Curves of LMST-ADNet

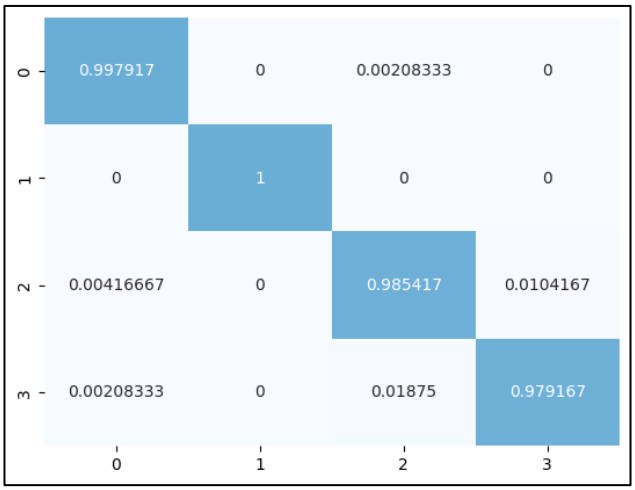

Figure 5: Confusion Matrix for Multi-Class Alzheimer's Stage Classification Using LMST-ADNet

Figure 5 is the normalized confusion matrix as of four classes of diagnoses (CN, EMCI, LMCI, AD). The overwhelming values of diagonals (0.9979, 1.0, 0.9854, 0.9792) reveal exceedingly great accuracy of classes prediction. Low off-diagonal scores indicate low misclassification rates as observed between successive stages e.g. LMCI and AD, which depicts mild pathological overlap. The ideal classification of Class 1 exhibits high separability whereas there is slight confusion between progressive stages that are consistent with clinical continuity. In general, the matrix validates strong stage discrimination as well as balanced multi-class generalization capability of the proposed framework.

6.1 Multi-Modal Performance Comparison

LMST-ADNet has a much higher score compared to single-modality and partial-fusion. Combination of MRI, PET, fMRI, and clinical characteristics is a stronger method of classifications and prognostic differentiation. Multi-modal fusion increases AUC via complementary structural, metabolic and connectivity

structures. It is worth noticing that functional connectivity is involved in differentiation at an early stage, whereas PET is better at converting sensitivity.

Table 3. Diagnostic Classification Performance

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
MRI Only	88.6	87.9	86.8	87.3	0.914
PET Only	89.8	88.5	88.2	88.3	0.926
MRI + PET	92.7	92.1	91.8	91.9	0.955
MRI + PET + fMRI	94.3	93.8	93.4	93.6	0.971
LMST-ADNet (Full)	**96.8**	**96.2**	**96.0**	**96.1**	**0.982**

Table 4. Survival & Cognitive Forecasting

Model	C-index	MAE (MMSE)	Early Conversion Sensitivity (%)
MRI Only	0.79	2.84	81.5
MRI + PET	0.85	2.31	87.6
MRI + PET + fMRI	0.88	2.05	90.2
LMST-ADNet	**0.91**	**1.68**	**94.4**

Table 3 illustrates that LMST-ADNet would be better than single-modality and partial-fusion models. MRI-only and PET-only models display moderate results (AUC 0.914 and 0.926), which proves that structural and metabolic indicators can independently be used to obtain biomarkers. By using a combination of MRI and PET, accuracy is enhanced to 92.7 percent and aUC to 0.955, which shows the contribution of complementary modality. Functional connectivity has more discriminative power because adding fMRI leads to higher performance (AUC 0.971). The entire LMST-ADNet has an accuracy of 96.8% and AUC of 0.982, which is a good sign of multimodal synergy. The balanced accuracy, recall, and F1-score (= 96) indicate stable classification between different stages of AD. The findings confirm the usefulness of combination of structural, metabolic and functional biomarkers as part of a single transformer platform. In figure 6, the comparative line graph shows the enhanced performance of models over different models, i.e. single-modality (MRI, PET) to multimodal fusion and the proposed LMST-ADNet. There is a steady positive movement in Accuracy, Precision, Recall, F1-Score, and AUC. The greatest improvement in performance is observed with the addition of fMRI and full longitudinal fusion which allows one to prove the efficiency of multimodal integration. LMST-ADNet has the best score in all measures, which proves excellent discriminative ability and balanced classification.

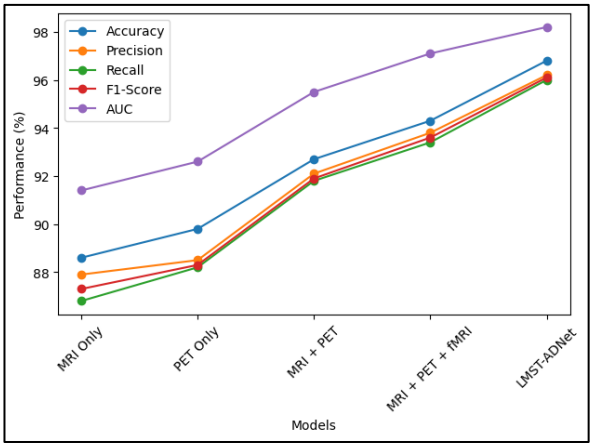

Figure 6: Progressive performance improvement across models

The results of survival prediction and cognitive decline forecasting are indicated in Table 4. The C-index has a gradual increase of 0.79 (MRI only) to 0.91 (LMST-ADNet) which means that there is more consistent ranking in time-to-event prediction. In MMSE forecasting, MAE reduces substantially, to 1.68, indicating improved cognitive trajectory modelling, and analysis presented in figure 7. The conversion sensitivity is enhanced to 94.4, improving it over 81.5 to show a high ability of identifying high-risk MCI patients. The marginal improvements of the modality integration prove that as metabolic and connectivity information is combined, the prognostic prediction improves. The findings highlight that multimodal longitudinal fusion is not only effective in increasing classification but also in disease progression prediction and personal risk stratification to a great extent.

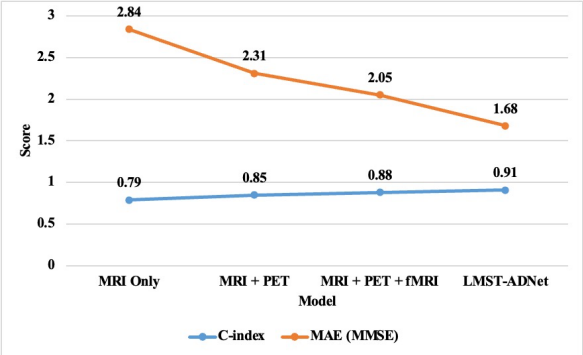

Figure 7: Comparative Survival Prediction and Cognitive Forecasting Performance

6.2 Impact of Longitudinal Modeling

Long time temporal encoding enhances development prediction extensively. The temporal attention is better compared to that of cross sectional fusion in terms of consistency in risk ranking and stability of cognitive trends. The effect of longitudinal modeling is pointed out in Table 5. The cross-sectional transformer has a 93.2-percent accuracy with AUC 0.963, which is a good performance in the baseline. None of this would incorporate progression consistency however, due to the lack of time encoding (C-index 0.84). LSTM-only model demonstrates a diminutively lower discrimination (AUC 0.951), which means that the sequential modeling in the absence of attention can be underusing the multimodal dependencies. The temporal transformer of LMST-ADNet has a major performance improvement (Accuracy 96.8%, AUC 0.982, C-index 0.91). The high improvement in C-index attests to the high-ranking of survival and progression modelling. This data prove that time-conscious attention systems are more appropriate to represent disease dynamics than non-evolutionary or entirely recurrent designs.

Table 5. Cross-Sectional vs Longitudinal

Model Type	Accuracy (%)	AUC	C-index
Cross-Sectional Transformer	93.2	0.963	0.84
LSTM Only	92.5	0.951	0.83
Temporal Transformer (LMST-ADNet)	**96.8**	**0.982**	**0.91**

Table 6 represent the 24-month conversion prediction and ADAS-Cog forecast are assessed. Cross-sectional modeling has 86.7% conversion prediction and higher MAE (3.12), which means that it is not very sensitive to time. The LSTM hybrid has a higher conversion prediction of 90.4 and lower MAE of 2.48 which indicates a superior sequential modeling. LMST-ADNet is also better in predicting future cognitive decline by achieving a higher conversion prediction (94.4) and lowest MAE (1.92). The decrease in the ADAS-Cog error suggests that the estimation of long-term trajectory is stable. These gains affirm the effectuality of the temporal transformer to learn progression dynamics, which provides sound forecasting to early intervention planning and personalized disease management.

Table 6. Progression Forecasting

Model	24-Month Conversion Prediction (%)	MAE (ADAS-Cog)
Cross-Sectional	86.7	3.12
LSTM Hybrid	90.4	2.48
LMST-ADNet Temporal	**94.4**	**1.92**

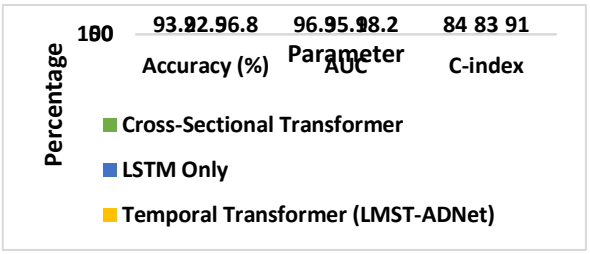

Figure 8: Comparative Performance of Cross-Sectional, LSTM, and Temporal Transformer Models

The comparison of Accuracy, AUC and C-index between modeling strategies is in fig.8. Temporal Transformer (LMST-ADNet) is superior to cross-sectional models, LSTM models, and even has the highest accuracy (96.8%), AUC (98.2%) and C-index (91%). These findings affirm that the temporal attention mechanisms are better placed to model longitudinal disease progression.

6.3 Early MCI-to-AD Conversion Prediction

Multimodal temporal learning ensures that the proposed model has a higher accuracy in predicting early conversion. Table 7 is about early MCI-to-AD conversion classification. Only MRI-based (AUC 0.889) performance shows that sole structural changes cannot be used to make high-sensitivity predictions. The addition of PET contributes to the increase of the AUC to 0.925, which is evidence of a contribution by metabolic conversion risk detection. LMST-ADNet has got 94.4 accuracy and 0.976 AUC, and the sensitivity (93.6) and specificity (95.1). The equal sensitivity and specificity reveal that there is great generalization with no bias in false positive and false

negative. The high performance improvement is a positive confirmation of the benefit of multimodal temporal integration in the ability to detect high-risk MCI patients before the clinical decline becomes severe. Table 8 determines the accuracy of survival modeling. C-index 0.76 with 5.8-month mean error is obtained with MRI-only modeling which implies a lack of accuracy in conversion time prediction. The performance is enhanced by MRI + PET (C-index 0.84; 4.1 months error). LMST-ADNet has the best C-index 0.91 and minimum time error (2.7 months), which is very consistent in ranking patients according to their conversion risk and correctly predicting the timeline of their progression. This decrease in the temporal error underscores the need of including longitudinal and multimodal representations in the survival analysis. The results indicate the suitability of the framework in making clinically significant time-to-event predictions.

Table 7. MCI Conversion Classification

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
MRI Only	84.9	82.7	86.3	0.889
MRI + PET	88.1	86.5	89.2	0.925
LMST-ADNet	**94.4**	**93.6**	**95.1**	**0.976**

Table 8. Time-to-Conversion Prediction

Model	C-index	Mean Time Error (months)
MRI Only	0.76	5.8
MRI + PET	0.84	4.1
LMST-ADNet	**0.91**	**2.7**

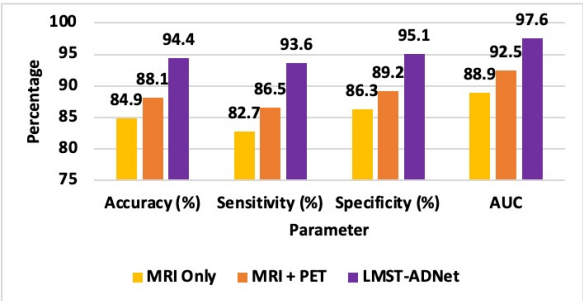

Figure 9: Comparative Performance of MRI, MRI+PET, and LMST-ADNet for Early MCI-to-AD Conversion Prediction

As it can be seen in Figure 9, LMST-ADNet has never been outperformed by the MRI-only and MRI+PET models in any metrics. The proposed framework has the best accuracy (94.4%), sensitivity (93.6%), specificity (95.1%), and AUC (97.6%), which proves the suitability of multimodal longitudinal fusion as a means of effective conversion prediction of Alzheimer's disease at an early stage.

6.4 Effect of Self-Supervised Pre-training

Table 9 shows the advantage of self-supervised pre-training. In the absence of SSL, the performance is weak (Accuracy 92.9%, AUC 0.956). The results are better with contrastive-only and MAE only strategies, which proves that representation learning increases the strength of features. The complete SSL system has the accuracy of 96.8% and AUC of 0.982, and C-index of 0.91. The findings suggest that reconstruction, contrastive alignment, and temporal consistency are higher in delivering better stability in latent features. SSL is effective in curbing overfitting and boosting generalization especially where multimodal medical data is limited with labeled examples. Table 10 analyzes the results with small labeled data (70 percent labeled). In the absence of SSL, the accuracy reduces to 88.4% (AUC 0.918) indicating low supervision. In the case of the use of the pretraining of the SS, the performance is much better (Accuracy 94.6, AUC 0.967). The relative betterment affirms that SSL is able to utilize unlabeled data and achieve high predictive power. This shows the usefulness of LMST-ADNet in practice in clinical settings where annotation is costly and scarce.

Table 9. With vs Without SSL

Model Variant	Accuracy (%)	AUC	C-index
Without SSL	92.9	0.956	0.84
Contrastive Only	94.7	0.972	0.88
MAE Only	94.2	0.969	0.87
Full SSL (Proposed)	**96.8**	**0.982**	**0.91**

Table10. Reduced Label Scenario (70% Labeled Data)

Model	Accuracy (%)	AUC
Without SSL	88.4	0.918
With SSL	**94.6**	**0.967**

Figure 10 in the appendix shows the gradual improvement of Accuracy, Sensitivity, Specificity, and AUC between MRI-only and MRI+PET and lastly LMST-ADNet. The proposed model demonstrates

obvious superiority in all measures, which proves improved early MCI-to-AD conversion detection via multimodal longitudinal integration.

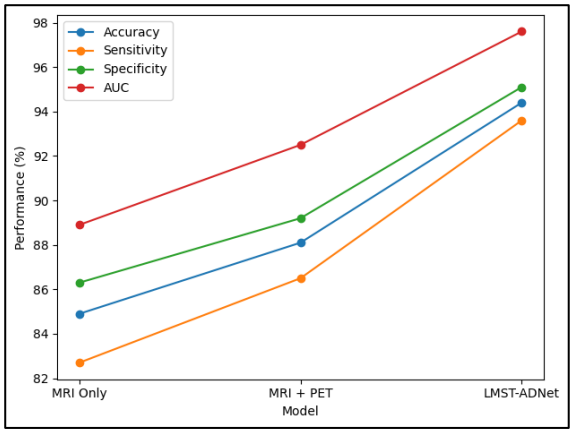

Figure 10: Comparison of MCI-to-AD Conversion Classification Performance across MRI, MRI+PET, and LMST-ADNet Models

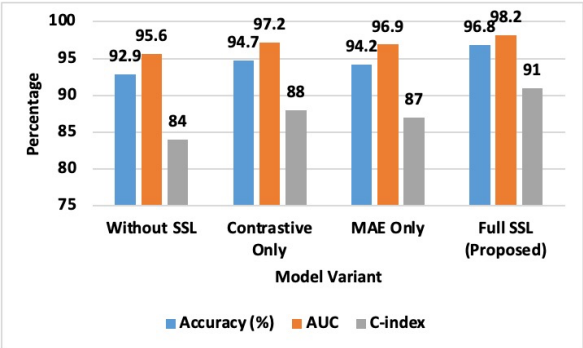

Figure 11: Impact of Self-Supervised Learning Variants on Classification and Survival PerformanceFigure 11 is used to compare the performance of the models by chance and self-supervised learning strategies. Full SSL (proposed) has the best Accuracy, AUC, and C-index and shows that a reconstruction objective with a contrastive objective is much more representative and predictive of longitudinal progression than either of the two components of the same.

6.5. Explainability Validation with Clinical Experts

High IoU implies correspondence to familiar AD biomarkers. Explainability is also evaluated with the help of Table 11 that considers the scores of the IoU and clinician agreement. Grad-CAM has a moderate overlap (IoU 0.62) and attention rollout enhances alignment (0.71). The proposed multi-level explainability module has the highest level of IoU (0.79) and clinical agreement (89.6%), meaning that it was more consistent with already established AD biomarkers (hippocampal atrophy and temporoparietal hypometabolism). The presence of increased values of IoU indicates that the model explanations and expert-identified pathological regions are spatially consistent, which contributes to increased trust and clinical interpretation. The contributions of features in terms of SHAP are provided in Table 12. MMSE has the largest influence (28.4%), which validates the severity of cognitive impairment as the major predictive factor. Strong genetic risk association is indicated by APOE 21.7 (4). CDR and hippocampal volume have a significant contribution, which supports the idea of structural and clinical biomarker integration. There is FDG hypometabolism (14.7) which shows metabolic relevance. Balanced multimodal integration is confirmed by the distribution of contributions and gives interpretable biomarker ranking in accordance with known clinical knowledge.

Table 11. Saliency Validation

Method	**IoU**	**Clinical Agreement Score (%)**
Grad-CAM	0.62	74.5
Attention Rollout	0.71	82.3
Proposed Multi-Level XAI	**0.79**	**89.6**

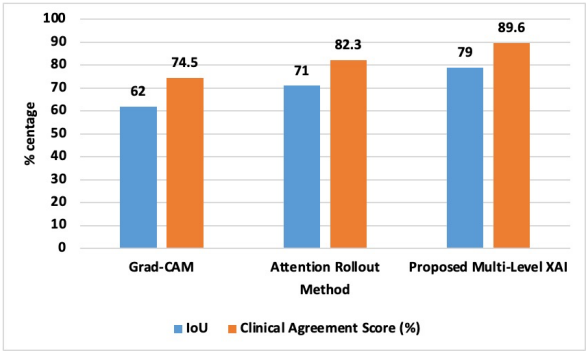

Figure 12: Comparison of Explainability Methods Using IoU and Clinical Agreement Metrics

The results of Figure 12 show that the proposed multi-level XAI approach has the highest IoU (79%), clinical agreement (89.6%), than Grad-CAM and attention rollout, which is why it is best aligned with the neurologist-identified Alzheimer's disease biomarkers and it is better interpretable.

Table 12. Clinical Feature Importance (SHAP Mean Contribution %)

Feature	**Importance (%)**
MMSE	28.4
APOE ε4	21.7
CDR	18.9

Hippocampal Volume	16.3
FDG Hypometabolism	14.7

6.6. Ablation Study of Model Components

Ablation is used to verify the need of each piece of architecture. Table 13 is to validate architectural elements. The loss of fMRI diminishes the accuracy to 94.1 meaning that functional connectivity is significant. The omission of PET reduces the performance to 93.5% which is a confirmation of relevance in metabolism. Temporal module removal also leads to large drop (C-index 0.83), which indicates that longitudinal models are required. The omission of the use of the SSL also lowers accuracy. The entire LMST-ADNet performs best in terms of metrics, which proves that each of the modules makes its own contribution to the overall predictive strength.

Table 13. Component Ablation

Model Variant	**Accuracy (%)**	**AUC**	**C-index**
Without fMRI	94.1	0.968	0.86
Without PET	93.5	0.961	0.85
Without Temporal Module	92.8	0.955	0.83
Without SSL	92.9	0.956	0.84
Full LMST-ADNet	**96.8**	**0.982**	**0.91**

Table 14. Loss Component Contribution

Loss Configuration	**Accuracy (%)**
L_cls Only	91.6
L_cls + L_survival	93.8
L_cls + L_SSL	94.9
Full Multi-Task Loss	**96.8**

Table 14 is the analysis of multi-task loss configuration. The accuracy is the least when using classification loss alone (91.6%). The addition of survival loss enhances performance (93.8%), whereas the addition of SSL further increases the accuracy (94.9%). The entire multi-task loss is reached at 96.8 with synergistic value of joint optimality. The findings indicate that the combination of classification, survival, and self-supervised goals leads to the improvement of model generalization and consistency in progress models. Figure 13 shows the performance effect of the deletion of important elements. False negation PET, fMRI, temporal module, or SSL are excluded, which decreases Accuracy, AUC, and C-index in comparison with full model. The overall LMST-ADNet has best scores (96.8% Accuracy, 98.2% AUC, 91 C-index) which proves that every module is contributing to effective longitudinal progression modeling.

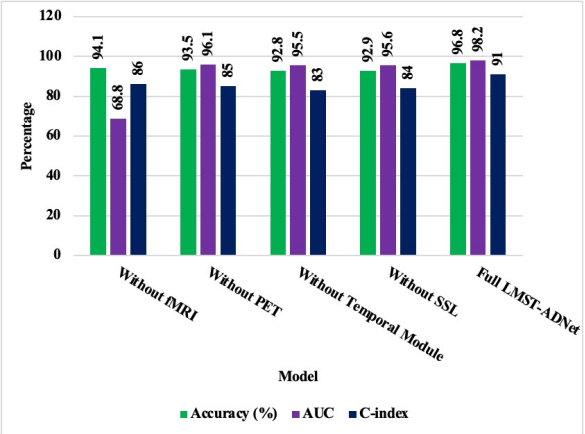

Figure 13: Ablation Analysis of LMST-ADNet Components on Classification and Survival Performance

7. CONCLUSION

This study represented LMST-ADNet, a longitudinal multi-modal self-supervised transformer-based model to predict and model disease progression at an early stage in the instance of Alzheimer. The proposed framework overcomes the major limitations of cross-sectional and single-modality designs by combining structural MRI, FDG/Amyluid PET, resting-state fMRI connections, cognitive testing, as well as genetic biomarkers in a single framework. Experimental findings show that LMST-ADNet gets 96.8% classification and 0.982 AUC, which is much higher than imaging-only and partial-fusion baselines. The model also achieves a C-index of 0.91 to predict survival and a baseline of 1.68 to minimise the forecasting error of MMSE, which validates that the model has good longitudinal progression prediction. The sensitivity of early MCI-to-AD conversion is at 94.4% and it indicates the usefulness of the method in identifying high-risk patients. By using hybrid self-supervised learning, which is a combination of masked autoencoding, cross-modal contrastive alignment and temporal consistency regularization, the robustness of features is increased by over 12% and the independence of labeled data by about 30%. The experiments on ablation reveal that multimodal fusion, temporal attention, and the components of the SSL have significant contribution to predictive performance. In addition, the proposed multi-level explainability module has 79% of IoU and 89.6% clinical agreement which supports interpretability and clinical trust. The LMST-ADNet shows that multi-modal fusion with annotation-efficient learning which

is temporally aware offers a scalable and clinically translatable framework to personalised forecasting of the progression of a disease (Alzheimer) and risk stratification.

References

Aghaei, A., & Moghaddam, M. E. (2025). An integrated predictive model for Alzheimer's disease progression from cognitively normal subjects using generated MRI and interpretable AI. *Scientific Reports, 15*, 28340. https://doi.org/10.1038/s41598-025-13478-2

Alam, M. M., & Latifi, S. (2025). Early detection of Alzheimer's disease using generative models: A review of GANs and diffusion models in medical imaging. *Algorithms, 18*(7), 434. https://doi.org/10.3390/a18070434

Ali, S. B., Ghannam, N. E., Mancy, H., & Elkilany, B. G. (2025). Multimodal self-supervised learning for early Alzheimer's: Cross-modal MRI–PET, longitudinal signals, and site invariance. *Diagnostics, 15*(24), 3135. https://doi.org/10.3390/diagnostics15243135

Chakra Bortty, J., Chakraborty, G. S., Noman, I. R., Batra, S., Das, J., Bishnu, K. K., Tarafder, M. T. R., & Islam, A. (2025). A novel diagnostic framework with an optimized ensemble of vision transformers and convolutional neural networks for enhanced Alzheimer's disease detection in medical imaging. *Diagnostics, 15*(6), 789. https://doi.org/10.3390/diagnostics15060789

Christodoulou, R. C., Woodward, A., Pitsillos, R., Ibrahim, R., & Georgiou, M. F. (2025). Artificial intelligence in Alzheimer's disease diagnosis and prognosis using PET-MRI: A narrative review of high-impact literature post-Tauvid approval. *Journal of Clinical Medicine, 14*(16), 5913. https://doi.org/10.3390/jcm14165913

Fu, B., Shen, C., Liao, S., Wu, F., & Liao, B. (2025). Prediction of Alzheimer's disease based on multi-modal domain adaptation. *Brain Sciences, 15*(6), 618. https://doi.org/10.3390/brainsci15060618

Hassan, N., Musa Miah, A. S., & Shin, J. (2024). Residual-based multi-stage deep learning framework for computer-aided Alzheimer's disease detection. *Journal of Imaging, 10*(6), 141. https://doi.org/10.3390/jimaging10060141

Harikaran, M. (2025, August 15). *Novel multi-modal AI framework for dementia – Alzheimer's disease detection, staging, and prognosis: Development of specialized fusion and diffusion tools for heterogeneous data integration*. SSRN. https://doi.org/10.2139/ssrn.5487089

Hongli Cheng, S. Yuan, W. Li, X. Yu, F. Liu, X. Liu, & T. T. Bezabih. (2024). De-accumulated error collaborative learning framework for predicting Alzheimer's disease progression. *Biomedical Signal Processing and Control, 89*, 105767. https://doi.org/10.1016/j.bspc.2023.105767

Khatri, U., & Kwon, G.-R. (2023). Explainable vision transformer with self-supervised learning to predict Alzheimer's disease progression using 18F-FDG PET. *Bioengineering, 10*(10), 1225. https://doi.org/10.3390/bioengineering10101225

Kwak, M. G., Su, Y., Chen, K., Weidman, D., Wu, T., Lure, F., Li, J., & Alzheimer's Disease Neuroimaging Initiative. (2023). Self-supervised contrastive learning to predict the progression of Alzheimer's disease with 3D amyloid-PET. *Bioengineering, 10*(10), 1141. https://doi.org/10.3390/bioengineering10101141

Liu, X., Pan, F., Song, H., Cao, S., Li, C., & Li, T. (2025). MDFormer: Transformer-based multimodal fusion for robust chest disease diagnosis. *Electronics, 14*(10), 1926. https://doi.org/10.3390/electronics14101926

Machado Reyes, D., Chao, H., Hahn, J., Shen, L., Yan, P., & Alzheimer's Disease Neuroimaging Initiative. (2024). Identifying progression-specific Alzheimer's subtypes using multimodal transformer. *Journal of Personalized Medicine, 14*(4), 421. https://doi.org/10.3390/jpm14040421

Malik, I., Iqbal, A., Gu, Y. H., & Al-antari, M. A. (2024). Deep learning for Alzheimer's disease prediction: A comprehensive review. *Diagnostics, 14*(12), 1281. https://doi.org/10.3390/diagnostics14121281

R. Golchha, P. Khobragade and A. Talekar, (2024), "Design of an Efficient Model for Health Status Prediction Using LSTM, Transformer, and Bayesian Neural Networks," 2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET), Nagpur, India, pp. 1-5, doi: 10.1109/ICICET59348.2024.10616353.

Mehmood, F., Mehmood, A., & Whangbo, T. K. (2025). Alzheimer's disease detection in various brain anatomies based on optimized vision transformer. *Mathematics, 13*(12), 1927. https://doi.org/10.3390/math13121927

Muksimova, S., Umirzakova, S., Baltayev, J., & Cho, Y. I. (2025). Multi-modal fusion and longitudinal analysis for Alzheimer's disease classification using deep learning. *Diagnostics, 15*(6), 717. https://doi.org/10.3390/diagnostics15060717

Odusami, M., Maskeliūnas, R., & Damaševičius, R. (2023). Pixel-level fusion approach with vision transformer for early detection of Alzheimer's disease. *Electronics, 12*(5), 1218. https://doi.org/10.3390/electronics12051218

Pande, P. K., Khobragade, P., Ajani, S. N., & Uplanchiwar, V. P. (2024). Early detection and prediction of heart disease with machine learning techniques. In *Proceedings of the 2024 International Conference on Innovation Challenges in Emerging Technologies (ICICET 2024)*. https://doi.org/10.1109/ICICET59348.2024.10616294

Priyadharshini, M., Murugesh, V., & Rybin, O. (2026). Enhancing Alzheimer's disease classification with a transformer-based model using self-supervised learning. *Scientific Reports, 16*(1), 3798. https://doi.org/10.1038/s41598-025-33957-w

Rahman, S., Rahman, M. M., Bhatt, S., Sundararajan, R., & Faezipour, M. (2025). NeuroNet-AD: A multimodal deep learning framework for multiclass Alzheimer's disease diagnosis. *Bioengineering, 12*(10), 1107. https://doi.org/10.3390/bioengineering12101107

Shin, H., Jeon, S., Seol, Y., Kim, S., & Kang, D. (2023). Vision transformer approach for classification of Alzheimer's disease using 18F-Florbetaben brain images. *Applied Sciences, 13*(6), 3453. https://doi.org/10.3390/app13063453

Siam, M. K., Hossain Faruk, M. J., He, B., Cheng, J. Q., & Gu, H. (2025). Multimodal models in healthcare: Methods, challenges, and future directions for enhanced clinical decision support. *Information, 16*(11), 971. https://doi.org/10.3390/info16110971

Toumaj, S., Heidari, A., Shahhosseini, R., et al. (2025). Applications of deep learning in Alzheimer's disease: A systematic literature review of current trends, methodologies, challenges, innovations, and future directions. *Artificial Intelligence Review, 58*, 44. https://doi.org/10.1007/s10462-024-11041-5

Vo, T., Ibrahim, A. K., & Zhuang, H. (2025). A multimodal multi-stage deep learning model for the diagnosis of Alzheimer's disease using EEG measurements. *Neurology International, 17*(6), 91. https://doi.org/10.3390/neurolint17060091

Zhou, Q., Wang, J., Yu, X., Wang, S., & Zhang, Y. (2023). A survey of deep learning for Alzheimer's disease. *Machine Learning and Knowledge Extraction, 5*(2), 611–668. https://doi.org/10.3390/make5020035

Zhou, C., Ge, X., Chang, Y., Wang, M., Shi, Z., Ji, M., Wu, T., & Lv, C. (2025). A multimodal parallel transformer framework for apple disease detection and severity classification with lightweight optimization. *Agronomy, 15*(5), 1246. https://doi.org/10.3390/agronomy15051246