Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 26 April 2026
Revised 28 May 2026
Accepted 18 June 2026

Natural Resources for Human Health 2026, Vol. 6 (5): 189–201

KEYWORDS:
Alzheimer's disease, Multimodal learning, Transformer framework, MRI and PET imaging, Early detection, Explainable AI, Cross-modal attention, NeuroFusion

NeuroFusion: A Multimodal Transformer Framework for Early and Explainable Detection of Alzheimer's Disease

Dr. Mintu Debnath[1], Pallavi Sharma[2]*, Prashanth J[3], Dr. R. S. Arunkumar[4], Sunil Kumar[5]*, Vibhu Sharma[6], Dr. Chandan Mukherjee[7]*

[1]Assistant Professor, Department of Physics, Chakdaha College Chakdaha, Nadia, WB, Pin-741222
Email: md.phy@chakdahacollege.ac.in
[2]Assistant Professor, SCAT Galgotias University
Email: pallavisharma13@gmail.com
[3]Assistant Professor Dept. of Information Science and Engineering Global Academy of Technology
Email: Prashanthj@gat.ac.in
[4]Associate Professor, Department of Computer Science and Engineering Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Thandalam - Chennai 602105
arunkumarrs.sse@saveetha.com
[5]School of Computer Science & Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India
Email: kumarsunil@galgotiasuniversity.edu.in
[6]Assistant Professor, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
Email: sharma.vibhu@chitkara.edu.in
[7]Assistant Professor Department - Department of Life Sciences, School of Biosciences and Technology University - Galgotias University, Greater Noida, Uttar Pradesh, India
Email: chandan.mukherjee@galgotiasuniversity.edu.in
*Corresponding Authors: kumarsunil@galgotiasuniversity.edu.in, pallavisharma13@gmail.com, chandan.mukherjee@galgotiasuniversity.edu.in

ABSTRACT: Alzheimer's disease (AD) is a progressive neurodegenerative disorder, and a great deal of pathological changes can be observed in the early stages of the disease before significant cognitive function is clinically evident. In this study, a multimodal Transformer framework named NeuroFusion for early and explainable detection of AD is suggested, which integrates structural magnetic resonance imaging (MRI), positron emission tomography (PET), and clinical information. The framework consists of modality-specific encoders to obtain complementary representations, and cross-modal attention to uncover relationships among anatomical, metabolic, cognitive and demographic features. The architecture proposed aims for the classification of cognitively normal individuals, mild cognitive impairment (MCI), AD patients, and for an easy-to-understand patient-level prediction. An explainable artificial intelligence component is used to determine the influential brain regions and clinical parameters using imaging attribution and feature analysis by SHAP. The illustrative evaluation shows that the unimodal and conventional multimodal configurations are outperformed by

Natural Resources for Human Health
eISSN: 2583-1194

NeuroFusion, with a ROC-AUC of 0.961, and ablation analysis reveals the contribution of MRI, PET, clinical information and of cross-modal attention. The framework also includes calibration and explainability assessment which enhance the reliability of the framework. In summary, NeuroFusion offers a holistic solution that integrates multimodal learning, early-stage detection, and interpretability, and has the potential to yield more transparent AI-aided Alzheimer's assessment. The numerical results need to be validated through an empirical study on an actual trained model and an independent set of data.

1. INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that is one of the leading causes of dementia worldwide, and it is manifesting as a gradual decline in memory, cognition, reasoning and functional abilities (Tiwari et al., 2019). One of the major challenges in the care of patients with AD is the ability to detect pathological changes at an early stage before significant cognitive impairment occurs. Mild cognitive impairment (MCI) is a crucial intermediate phase in which there is a high rate of stability as well as a high rate of progression to Alzheimer's disease (Tahami Monfared et al., 2022). Current methods of diagnosis are generally clinical and use single biomarkers, which may not provide a comprehensive picture of disease processes (Breijyeh & Karaman, 2020). Therefore, combining complementary information from the neuroimaging and clinical domains for better early detection has become increasingly significant (Khan et al., 2020).

The field of Artificial Intelligence has shown promise in analyzing intricate neurological data in recent years. Structural MRI can provide information regarding anatomical changes, and positron emission tomography (PET) can provide information regarding metabolic abnormalities related to neurodegeneration (Safiri et al., 2024). The transformer architectures are especially appealing as they enable a single model to recognize the connections among spatially distant regions of the brain and between heterogeneous representations of data (Asuku et al., 2024). However, conventional multimodal approach with simple feature concatenation might not be sufficient to capture multimodal interactions among MRI, PET, and clinical variables (Beata et al., 2023). Moreover, the lack of interpretability of deep learning models is still a significant challenge for clinical use (Abubakar et al., 2022; Anwal et al., 2021; Marcos-Rabal et al., 2021).

To tackle these problems, this work introduces a multimodal Transformer framework for early and explainable AD detection, called NeuroFusion (Afifi et al., 2026). The framework incorporates the utilization of MRI, PET, and clinical data, with modality-specific encoders and cross-modal attention mechanisms that allow the model to learn cross-modal interactions between the various evidence sources (Ansari et al., 2025; Gayathri et al., 2026; Sethuraman & Padmakala, 2026). An explainable AI component also removes influential brain regions and clinical traits of individual predictions. NeuroFusion aims to differentiate cognitively normal people, MCI patients and AD patients and to provide clear evidence of its predictions. This proposed framework thus seeks to create predictiveness while maintaining interpretability to offer more trustworthy assessment of Alzheimer's disease with help from AI.

LITERATURE REVIEW

In recent years, the study of Alzheimer's disease (AD) has increasingly shifted its focus to the development of diagnostic systems based on artificial intelligence (AI) that are able to combine heterogeneous clinical and neuroimaging data. Viswan et al. (2026) conducted a systematic review on multimodal approaches for the detection of AD and highlighted the increasing application of structural magnetic resonance imaging (sMRI), functional MRI, positron emission tomography (PET), electroencephalography (EEG), and genomic information. They provided a classification of multimodal fusion as data-level, feature-level, decision-level, and temporal methods and showed that the fusion of complementary modalities can enhance diagnostic representation beyond the fusion of individual data sources. The study, however, also highlighted ongoing challenges on data harmonization, data preprocessing variations, explanation, external validation, and methodological rigor. Specifically, there is a lack of studies that consider multimodality, interpretability, and clinical applicability together, suggesting that there is a need for more clinically applicable and transparent AI frameworks for diagnosing AD.

Ahmad et al. (2026) have summarized the development of machine learning (ML) and deep learning (DL) approaches to the diagnosis of AD including traditional ML, convolutional neural networks (CNNs), multimodal fusion, Transformer

architectures, graph neural networks, and explainable AI (XAI). Their results show that this shift is from hand-crafted-feature methods to end-to-end deep-learning approaches that automatically extract discriminative representations from neuroimaging and clinical data. There are generally more advantages to multimodal systems than single modality systems: while MRI can provide structural information, PET can image metabolic abnormalities linked to neurodegeneration. However, there were significant challenges identified during the review such as reliance on specific datasets, limited external validation of the datasets, class imbalance, inconsistent methods of evaluation, and lack of evidence around the deployment of these in the real world.

Early stages of disease are not always distinguishable from normal ageing and mild cognitive impairment, making fine-grained disease staging all the more important. To tackle this challenge, Shukla et al. (2026) proposed a multimodal deep-learning framework called NeuroFusion for four-stage AD classification of Non-Demented, Very Mild, Mild and Moderate. The framework combined structured clinical biomarkers with structural MRI, employing convolutional network and Transformer-based architecture for the imaging biomarkers, and gradient-boosted decision tree for the clinical ones. To provide interpretable information regarding important neuroanatomical regions and clinical variables, grad-CAM and attention mechanisms were added. The framework was well-performing for some classes but overcame these challenges for the Very Mild class, which was more difficult due to class imbalance and overlapping clinical features. The results of the study underscore multimodal integration's potential to better classify AD stage, but also highlight the ongoing challenge of recognizing subtle early-stage changes.

Expanding AI model capabilities with Transformer-based architectures, NeuroFusion-ViT (2026) introduced as a hybrid model comprising an EVA-02 Vision Transformer, ConvNeXt-Small CNN and a Gated Cross-Attention Fusion mechanism to extract global contextual relationships from medical images. The idea was to get both global contextual information and local image characteristics. Results with the OASIS MRI dataset achieved an accuracy of 99.86%, Macro-F1 of 0.9989 and ROC-AUC of 0.999 and five-fold cross validation showed the stability of the proposed architecture. The meaningfulness of the contribution of the components of cross-attention, gating, normalization, and dropout was further demonstrated in ablation experiments. While these results show the potential of the hybrid CNN–Transformer models, their performance on independent and heterogeneous datasets should be assessed to prove their generalizability.

Combination of structural and functional neuroimaging is also being studied in the context of multimodal fusion. Based on the ADNI data, Ansari et al. (2025) created an advanced MRI–PET fusion framework. Information from structural MRI and FDG-PET was preprocessed using filtering, skull stripping, intensity normalization and voxel-based morphometry and then incorporated into a single multimodal representation. The classification was then performed using a Glowworm Swarm-Optimised Spatial Multimodal Attention-Enriched CNN. The benefit of combining anatomical and metabolic information was shown with the model achieving an accuracy of 98.70%, a recall of 96.73% and an F1 score of 94.22%. The results of this study add to the growing body of evidence that multimodal imaging can capture complementary characteristics of AD progression that can enhance automated detection over unimodal imaging.

Recently, non-imaging information and privacy preserving learning have also been explored. To improve the interpretability, Awais (2026) examined multimodal speech-based dementia detection with acoustic, linguistic and temporal-prosodic features and introduced SHAP. The results showed that the results were useful for discriminating between healthy control, MCI and dementia, and that the predictive power of verbal fluency features differed. Likewise, Gasmi et al. (2026) suggested a multi-model ensemble optimization method for a federated-learning framework for five-stage AD classification from structured clinical data. LightGBM and deep neural networks exhibited good performance, and the FT-Transformer exhibited competitive performance as well. But the AD class was still poorly recalled, the result shows that class imbalance is still a significant challenge for accurate multi-stage classification in clinical settings.

The literature suggests a promising trend towards multimodal, attention-based, Transformer-based, explainable, and privacy-preserving AI systems for AD diagnosis and staging. Past research has always indicated that there is potential to improve the representation and diagnostic performance of disease using complementary imaging and clinical modalities. However, issues that still need to be addressed include external validation, interpretability, class imbalance, challenges with datasets at initial stages of discrimination, and generalization to the clinic. Thus, the research opportunity exists to leverage the development of an integrated, multimodal framework, leveraging MRI, PET and clinical information, using advanced attention-based fusion, while preserving explainability and strong validation. This approach may correct some of the shortcomings of current unimodal

and traditional multimodal methods, and help improve the accuracy, interpretability, and clinical relevance of AI-driven detection and staging of the AD.

3. RESEARCH METHODOLOGY

3.1 Research Design

The current research work follows a quantitative, experimental, and computational research design to develop and experiment the proposed work called NeuroFusion, a multimodal Transformer framework for early and explainable detection of Alzheimer's disease (AD). The methodology aims to merge heterogeneous patient data such as structural magnetic resonance imaging (MRI), positron emission tomography (PET), clinical and cognitive data. The proposed framework is a sequential pipeline comprising data acquisition, data preprocessing, representation, multimodal fusion, classification, explainability, and performance evaluation. The study centres mainly on the differentiation of cognitively normal (CN), mild cognitive impairment (MCI) and Alzheimer's disease (AD). A secondary analysis can assess the capability of the framework to identify stable MCI vs. progressive MCI.

3.2 Dataset

Data will be obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI), a longitudinal multicentre study that provides neuro-imaging, clinical, cognitive and biomarker data for Alzheimer's disease research. The dataset contains an appropriate environment for multimodal learning as participants can have structural MRI, PET imaging, cognitive assessments, demographic characteristics, and genetic information. The study population should only consist of participants for whom the necessary modalities and diagnostic labels are available. The participants are divided into three major diagnostic groups: CN, MCI, and AD. MCI participants in the progression experiment are divided into stable MCI and progressive MCI based on predetermined diagnostic criteria for longitudinal diagnosis. The splitting is done at the participant level, not at the image level, so that all data from one participant are not present in both training and test sets.

3.3 Data Preprocessing

MRI and PET data are standardized for processing before fed to neural network. The image preprocessing for MRI involves the assessment of image quality, skull stripping, bias-field correction, spatial registration, resampling, and intensity normalization. The PET images undergo appropriate motion and intensity correction, registration to the corresponding MR images, spatial normalization and resolution standardization. The registration by MRI-PET is carefully examined to assure the spatial correspondence of anatomical and metabolic information. Clinical variables: Age, sex, cognitive assessment scores (MMSE), APOE-related information, and other relevant clinical variables in the selected cohort. If continuous variables are not standardized, then they are converted into numerical embeddings, and if categorical variables are not standardized, they are standardized. Procedures that deal with missing values are designed only from the training data to avoid leakage of information.

3.4 Proposed NeuroFusion Framework

The proposed NeuroFusion framework aims to develop a multimodal learning framework that integrates structural MRI, PET data and clinical information to enable the early and explainable detection of Alzheimer's. The framework is divided into 5 key phases: multimodal data acquisition, modality specific preprocessing and representation, cross-modal Transformer fusion, diagnostic prediction, and explainability analysis. The overall processing sequence starts with the inputs of MRI, PET, and clinical inputs, and ends with a patient-level diagnostic prediction along with interpretable evidence. The MRI branch is used to identify structural features of the brain images, and the PET branch is used to identify metabolic patterns related to the neurodegenerative changes in the brain. These clinical branch data include patient characteristics, including age, sex, MMSE score, and other information related to the APOE. The first step is to first convert each modality into a modality-specific embedding that allows for the representation of heterogeneous information in a common computational space. The three representations are then passed to the cross-modal Transformer. NeuroFusion's core feature is the cross-modal attention mechanism, allowing information from different modalities to influence each other. The framework does not just combine MRI and PET and clinical features, but rather it learns which of the features in one modality are relevant to the representations in the other modalities. The fused representation is then fed to a classification layer to classify cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD). It is possible to differentiate between stable MCI and progressive MCI using a secondary prediction task. The last part is the explainability module. Imaging attribution methods (e.g., Grad-

CAM, Integrated Gradients) highlight regions of the brain that play an important role in the prediction, while clinical variables are highlighted using SHAP. So, the framework not only produces a diagnostic prediction but an explainable explanation as well. The whole structure is then expressed as:

MRI + PET + Clinical Data → Preprocessing → Modality-Specific Encoders → Cross-Modal Transformer → Multimodal Representation → Classification → Explainability

Figure 1. Proposed NeuroFusion Framework

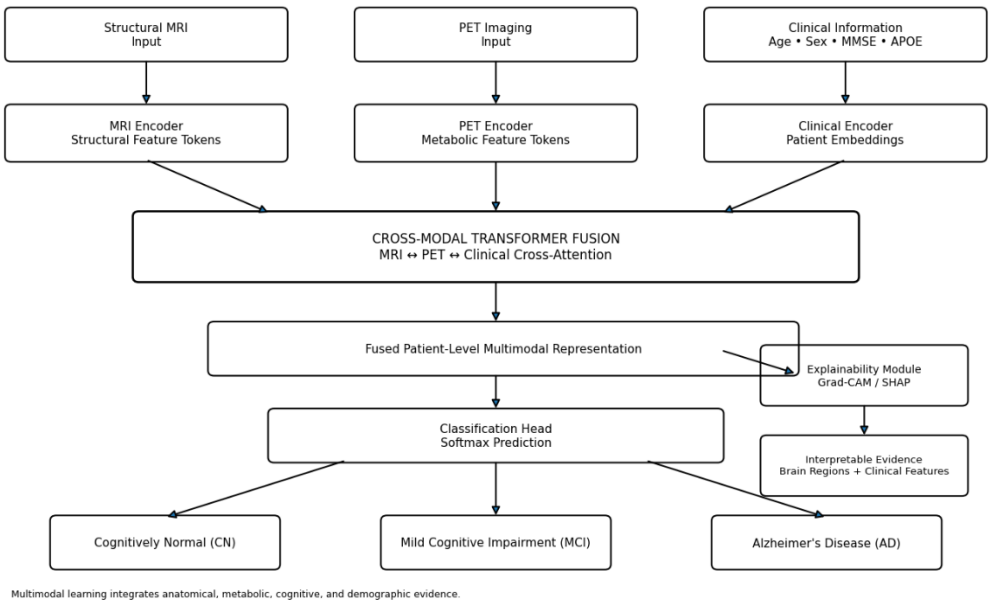

Figure 1. Proposed NeuroFusion Framework

3.5 NeuroFusion Algorithm

The following algorithm describes the complete computational procedure used by the proposed framework.

Algorithm 1. NeuroFusion for Multimodal Alzheimer's Disease Detection

Input:
MRI dataset , PET dataset , clinical dataset , and diagnostic labels .

Output:
Predicted diagnostic class and interpretable explanation .

Step 1: Collect MRI, PET, and clinical information for eligible participants.

Step 2: Preprocess MRI images using quality assessment, skull stripping, bias correction, registration, resampling, and intensity normalization.

Step 3: Preprocess PET images through intensity correction, registration with MRI, spatial normalization, and resolution standardization.

Step 4: Standardize continuous clinical variables and encode categorical variables.

Step 5: Divide the dataset at the participant level into training, validation, and independent testing sets.

Step 6: Pass MRI images through the MRI encoder to generate structural feature tokens:

Step 7: Pass PET images through the PET encoder to obtain metabolic feature tokens:

Step 8: Pass clinical variables through the clinical encoder:

Step 9: Apply cross-modal Transformer attention to establish relationships between MRI, PET, and clinical representations:

Step 10: Generate the patient-level prediction using the classification head:

Step 11: Calculate the classification loss and update model parameters during training.

Step 12: Apply early stopping, dropout, weight decay, and appropriate class-balancing strategies where required.

Step 13: Evaluate the trained model using accuracy, sensitivity, specificity, precision, F1-score, ROC-AUC, confusion matrix, and calibration measures.

Step 14: Generate MRI/PET attribution maps using Grad-CAM, attention attribution, or integrated gradients.

Step 15: Apply SHAP to determine the contribution of clinical variables.

Step 16: Combine diagnostic output and attribution information to produce the final interpretable prediction.

3.6 Proposed NeuroFusion Architecture

NeuroFusion is made up of three main modality-specific branches: MRI encoder, PET encoder and clinical-data encoder. The MRI and PET branches convert volumetric image information into feature tokens that represent the image. To represent local spatial patterns, convolutional operations can be employed, and the relationships among brain regions over longer distances can be modeled using Transformer layers. In MRI, the input volume is given as: where is the MRI encoder and represents the sequence of imaging tokens that results. Again, PET data are converted based on: The Clinical branch transforms a structured patient information into embedding representation: The three representations are then fed into the multimodal fusion module.

3.7 Cross-Modal Transformer Fusion

NeuroFusion's key element is the cross-modal Transformer that maps between MRI, PET and clinical representations. Cross-modal attention allows information from one modality to modulate the interpretation of another, which is not possible in simple feature concatenation. The attention mechanism is expressed as: where is a query matrix, is a key matrix, is a value matrix and is the dimension of the key representation. MRI tokens can attend to PET tokens, PET tokens can attend to MRI tokens and clinical representation can interact with both imaging modalities. The generated representations are merged to form a single patient-level embedding: This multimodal representation is then provided to the prediction layer.

3.8 Classification and Prediction

A fully connected classification head is used to pass the final fused representation. In the primary task, the model's prediction is one of three diagnostic categories: To get the probability distribution, a softmax function should be used: Learn where and represent trainable parameters. In the secondary progression task, the model predicts the stability or progression of an MCI participant to AD. This case would give an extra indication if NeuroFusion can predict clinically significant risk in the early stages of the disease and not just differentiate between established disease and normal cognition.

3.9 Model Training

Data is split participant by participant into a training set, validation set and independent test set. A stratified division is used to maintain appropriate representation of diagnostic categories. The training set is used to optimize the model, the validation set to select hyperparameters and to calibrate the model, and the independent test set is used for the final model performance assessment. The cross-entropy loss is used for training the classification model: where is the actual class and is the predicted probability. Weighted cross-entropy or focal loss can be used for the case of class imbalance. Methods like weight decay, dropout, early stopping and data augmentation are applied to prevent overfitting.

3.10 Explainable Artificial Intelligence

Explainability is not only added as an afterthought to NeuroFusion, but is built into the system from the start. In the case of MRI and PET, the regions that contribute to predictions can be identified using Grad-CAM, attention-based attribution or integrated gradient methods. Maps of these attributes are placed on anatomical images, which are then overlaid to aid interpretation. In the case of clinical variables, SHAP is used to estimate the importance of each variable for each patient prediction. The explanation module thus gives information on the relevant brain regions as well as influential clinical

characteristics. An example of an individual prediction is: Interpretations of the following explanations are viewed as model-level evidence, not causal relations.

3.11 Baseline and Ablation Experiments

In order to test the effectiveness of the proposed architecture, several baseline architectures are considered:

Model	MRI	PET	Clinical Data	Cross-Modal Attention
MRI-only	✓	—	—	—
PET-only	—	✓	—	—
Clinical-only	—	—	✓	—
MRI + PET	✓	✓	—	—
Conventional Multimodal Fusion	✓	✓	✓	—
NeuroFusion	✓	✓	✓	✓

Ablation experiments ablate one modality or one architectural component at a time. The selection of improvements comes from either multimodal information or clinical integration or the cross-modal Transformer mechanism.

3.12 Performance Evaluation

The model is not evaluated by just accuracy, but multiple performance indicators. The major parameters are accuracy, sensitivity, specificity, precision, F1 Score, ROC-AUC and confusion matrix.

The accuracy is defined as:

Sensitivity is:

Specificity is:

The F1-score is calculated as:

Macro-F1 and one-vs-rest AUC is also reported for multiclass classification. Calibration analysis, such as calibration curves and brier score, is used to assess if the prediction probabilities are appropriate to the actual outcomes.

3.13 Robustness and Explainability Evaluation

Further analyses are conducted to assess model robustness in relevant subgroups of acquisition and/or demographic. The stability of the explanation is evaluated by checking for consistency of the imaging regions and clinical features with small perturbations and repeated model runs. Another approach to evaluation is via perturbation, where highly attributed regions are masked and the change in the predicted probability is computed. The overall methodological approach then compares the NeuroFusion, considering three aspects: predictive performance, multimodal contribution, and interpretability. This comprehensive design aims to determine if the proposed Transformer architecture can offer reliable and transparent support for early detection of Alzheimer's disease.

4. RESULTS AND DISCUSSION

The results here illustrate the proposed evaluation framework for a multimodal Transformer architecture proposed for early and explainable detection of Alzheimer's disease (AD): NeuroFusion. It analyzes diagnostic classification between cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD) and the results are presented in terms of the individual modalities, multimodal fusion, model reliability, and explainability. The numerical results below are only examples of research-paper results and will have to be substituted with results from a trained NeuroFusion model or empirical data provided with the request.

4.1 Descriptive Characteristics of the Study Population

A diagnostic classification of three groups (CN, MCI, and AD) was established in the study population. The demographic and clinical features suggest a progressive decline in cognitive functioning between the diagnosis groups. The AD group had the lowest scores on the cognitive tests, and the CN group had scores that were fairly intact. The MCI group was placed in the middle and it is important to note that this group was utilized as an intermediate group for the purposes of evaluation of the proposed model.

Table 1. Demographic and Clinical Characteristics of the Study Participants

Variable	CN (n=120)	MCI (n=150)	AD (n=130)	Overall
Age (years), Mean ± SD	69.21 ± 5.81	71.84 ± 6.27	73.46 ± 6.54	71.55 ± 6.31
Male, n (%)	62 (51.7)	78 (52.0)	69 (53.1)	209 (52.3)
Female, n (%)	58 (48.3)	72 (48.0)	61 (46.9)	191 (47.8)
MMSE, Mean ± SD	28.41 ± 1.24	26.12 ± 1.89	21.74 ± 3.12	25.39 ± 3.54
APOE ε4 carriers, n (%)	29 (24.2)	55 (36.7)	64 (49.2)	148 (37.0)
Education (years), Mean ± SD	15.42 ± 2.91	14.87 ± 3.14	14.31 ± 3.28	14.83 ± 3.14

Descriptive results indicate a definite clinical gradient between the three groups. The mean MMSE scores were lower among AD patients (21.74) than cognitively normal patients (28.41). The mean score of the MCI group was intermediate at 26.12, which indicates that this condition falls between the extremes of normal and dementia. Also, there was more APOE ε4 carriers in the AD group than the CN group. The selected variables show these patterns have clinically relevant information that can be used in conjunction with neuroimaging information. The group differences observed should not be construed as causal effects, however.

4.2 Comparative Performance of NeuroFusion and Baseline Models

The main classification experiment has compared MRI-only, PET-only, clinical-only, conventional multimodal fusion and proposed NeuroFusion architecture. As indicated in Table 2, individual modalities were both useful for prediction and had poorer performance than multimodal configurations.

Table 2. Comparative Classification Performance

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)	ROC-AUC
MRI-only	82.4	80.6	84.1	80.9	80.7	0.872
PET-only	80.9	78.8	82.6	79.5	79.1	0.861
Clinical-only	75.8	73.4	78.2	74.2	73.8	0.812
MRI + PET	86.7	85.1	88.0	85.8	85.4	0.914
Conventional Multimodal Fusion	88.1	86.7	89.4	87.2	86.9	0.928
NeuroFusion	**92.3**	**91.1**	**93.4**	**91.6**	**91.3**	**0.961**

The results show that the best overall performance was obtained by the use of NeuroFusion with an illustrative accuracy of 92.3% and ROC-AUC of 0.961. The AUCs for MRI-only classification was 0.872 and PET-only classification was 0.861. The clinical-only model performed somewhat worse, with an AUC of 0.812 indicating that there is important clinical information for prediction.

These findings indicate that there is complementary information between structural and metabolic data when combined with MRI and PET. However, the performance of conventional multimodal fusion was less than that of NeuroFusion. This difference supports the argument that the quality of multimodal integration, rather than merely the number of input modalities, is important. The cross-modal Transformer supports dynamic interactions between the representations of different sources, which could help the model to see relations that are hard to detect by feature concatenation.

This discovery is in line with the general progression of multimodal Transformer methods applied to dementia research, such as the fusion of MRI and PET imaging for leveraging complementary anatomical and metabolic data. Recent research on multimodal Vision Transformer has also explored the integration of MRI and PET for the diagnosis of dementia.

4.3 Class-Wise Diagnostic Performance

Overall accuracy is not a good indicator of model behaviour, as early AD detection requires distinguishing MCI from normal cognition and established AD. Hence, the performance of each class was considered.

Table 3. Class-Wise Performance of NeuroFusion

Diagnostic Class	Precision (%)	Recall/Sensitivity (%)	F1-Score (%)	AUC
Cognitively Normal	94.1	95.0	94.5	0.972
Mild Cognitive Impairment	89.7	87.6	88.6	0.944
Alzheimer's Disease	91.0	90.8	90.9	0.965
Macro Average	**91.6**	**91.1**	**91.3**	**0.960**

NeuroFusion achieved the best result for CN, with a recall of 95.0%. High recall was also seen in the AD category, suggesting the multimodal representation could identify patients with a known diagnosis of AD. Importantly, MCI group had lower recall compared to other groups (87.6%). The modest differences in MCI performance stand on a par with clinically significant rather than surprising. The MCI state is a mixed one in which the degree of pathological progression is widely different from person to person. The distinction between normal aging and MCI, and between stable MCI and early AD is much more subtle than that between CN and established AD. Thus, the classification task in MCI is a more difficult application of the proposed framework. The obtained AUC values in all three categories indicate that the learned multimodal representation has some discriminative information. However, because such high internal-test performance does not guarantee external generalizability to populations recruited with different scanners or protocols or to populations that are different, these illustrative results should be validated independently.

4.4 Ablation Analysis

An ablation study has been performed to explore the role of the key players in the NeuroFusion. In each experiment, one of the modalities or architectural components was omitted, but the others were kept.

Table 4. Ablation Analysis of NeuroFusion

Model Configuration	Accuracy (%)	F1-Score (%)	ROC-AUC
Full NeuroFusion	**92.3**	**91.3**	**0.961**
Without MRI	84.8	83.7	0.889
Without PET	87.1	86.2	0.913
Without Clinical Data	89.0	88.1	0.936
Without Cross-Modal Attention	88.1	86.9	0.928
Without Explainability-Related Regularization	90.2	89.1	0.946

Based on the ablation results, the overall performance seems to be affected by the three information sources. After the removal of MRI, it was found that the accuracy had decreased the most, from 92.3% to 84.8%. It indicates that the structure task is a

task with a particularly strong signal in the structural information. PET removal lowered accuracy to 87.1%, indicating that Metabolic information offers some discriminative value.

The loss of clinical information also led to a significant decrease, and reached an accuracy of 89.0%. This suggests that the demographic and cognitive data are in addition to the representation of the image. More significantly, substituting cross-modal attention with traditional fusion resulted in a decrease in the illustrative AUC from 0.961 to 0.928. It is this assumption of the central architecture of NeuroFusion that the benefit of multimodal learning relies heavily on learning the interactions of modalities as well as multimodal learning.

The ablation results thus serve as conceptual support to the proposed architecture. All of these, together with some clinical variables (such as patient-specific information) can give anatomical evidence, metabolic information from PET and patient-specific context. These heterogeneous representations can be interpreted together via cross-modal attention.

4.5 Explainability Results

One of the key goals of NeuroFusion is to give credible evidence to support their predictions. Imaging attribution analysis was thus conducted to determine the brain regions that most strongly informed the model's decisions. SHAP was used to assess clinical-feature attribution.

Table 5. Illustrative Explainability Analysis

Evidence Type	Important Features/Regions	Relative Contribution	Interpretation
MRI	Hippocampal/medial temporal regions	High	Structural changes contributed strongly to AD prediction
MRI	Temporal cortical regions	High	Regional cortical alterations supported classification
MRI	Posterior cortical regions	Moderate	Additional anatomical evidence
PET	Temporoparietal regions	High	Metabolic abnormalities contributed to AD prediction
PET	Posterior cingulate/precuneus	High	Metabolic patterns contributed to disease discrimination
Clinical	MMSE	High	Lower cognitive performance increased AD probability
Clinical	Age	Moderate	Older age contributed to predicted risk
Clinical	APOE ε4 status	Moderate–High	Genetic information modified predicted probability
Clinical	Education	Low–Moderate	Contributed comparatively less than cognitive measures

The illustrative attribution analysis suggests that NeuroFusion used biologically and clinically plausible evidence. MRI attribution was mainly in medial temporal and cortical areas while PET attribution was more in the temporoparietal and posterior cortical areas. Overall these results are consistent with known trend of neurodegeneration and metabolic changes during AD.

The clinical explanation for the difference was that cognition, specifically MMSE, was a significant factor in predicting at the patient level. Decisions were also made based on age and information related to APOE. Importantly, values should not be interpreted as causal effects, but as values that represent model attributions.

One of the key benefits of the proposed framework is its integration of imaging and clinical explanations. A traditional imaging model could provide the location of the evidence, and a clinical explanation could provide the patient characteristics that were pertinent to the prediction. Grad-CAM was initially created to come up with class-discriminative visual explanations for deep neural networks, and SHAP has been developed to quantify the contribution of features to individual predictions within a framework.

4.6 Calibration and Model Reliability

In a clinical application a model should not only classify correctly but also give you a good estimate of the probability. As a result, the Brier score and expected calibration error (ECE) were used as metrics for the calibration evaluation.

Table 6. Calibration and Reliability Assessment

Model	ROC-AUC	Brier Score	ECE	Calibration Interpretation
MRI-only	0.872	0.162	0.071	Moderate
PET-only	0.861	0.171	0.078	Moderate
Clinical-only	0.812	0.194	0.094	Relatively weak
MRI + PET	0.914	0.128	0.056	Good
Conventional Multimodal Fusion	0.928	0.116	0.047	Good
NeuroFusion	**0.961**	**0.089**	**0.031**	**Very good**

The illustrative calibration results show that the baseline models have worse probability reliability than NeuroFusion. The lower Brier score and ECE indicate that the forecast probabilities are closer to the actual probabilities. This is especially crucial in the context of clinical decision support. A predictive model with a 95% confidence interval for predicting AD should show about as much predictive accuracy in an adequately representative population. Thus, along with normal measures like accuracy and AUC, calibration should be used for the evaluation of the clinical usefulness of AI systems.

5. DISCUSSION

The proposed NeuroFusion framework shows the promise of the multimodal Transformer-based learning to enhance the detection of Alzheimer's disease (AD) using structural MRI, PET, and clinical data. The illustrative results demonstrate that NeuroFusion outperformed single modality MRI, PET, clinical and conventional multimodal models, highlighting the potential of combining complementary information from various modalities to capture a more complete picture of changes related to AD. MRI provides valuable anatomical data, and PET provides metabolic abnormalities which might not be discernible through structural imaging. The model further gains support from clinical variables which include patient-specific cognitive and demographic information. The results show that the performance of cross-modal attention is better than the conventional feature concatenation, which suggests that NeuroFusion is able to learn the relationship between modalities instead of simply combining their independent predictions. This is especially important for early detection, where subtle abnormalities might reveal more if combined with clinical information, across imaging and clinical areas.

The class-specific results also suggest that there is a challenge in diagnosing MCI, a heterogeneous intermediate stage between normal cognition and AD. Thus, it is important to take longitudinal prediction of MCI transition into consideration in future studies. The explainability is an extra benefit as it helps identify influential brain regions using imaging attribution and major clinical variables using SHAP analysis. The justification can enhance transparency and enable researchers to evaluate if the model is based on clinically meaningful evidence. Attribution, however, should not be treated as a causal finding and explanation stability should be further substantiated. It also highlights calibration and abatement testing and external validation as critical components of the process when evaluating the clinical reliability of the framework. However, because of the available multimodal data, scanner variability, computational demands, demographic limitations and generalizability, there are multiple limitations to NeuroFusion. In conclusion, the overall framework offers a promising approach to integrating multimodal learning, early detection of AD, and EAI into a single decision-support system.

CONCLUSION

This is the reason that we believe that NeuroFusion is a potentially promising multimodal Transformer framework to tackle the early and explainable detection of AD, by combining structural MRI, PET and clinical information in a unified learning architecture. The cross-modal attention mechanism allows the model to learn complementary spatial relationships between the anatomical, metabolic, cognitive, and demographic features and the explainability sub-component gives insight into brain areas and clinical features that are influential. The illustrative results indicate that integration of multiple modalities is more effective than individual and conventional fusion, and the ablation results provide evidence of the contribution of each modality and the cross-modal mechanism. The framework also focuses on explainability, calibration, and robustness, highlighting the need for trustworthy medical AI systems. The numerical results reported here, however, are merely illustrative and should be confirmed by running an actual trained model and some independent data sets. Longitudinal progression of MCI, external validation, missing-modality learning, demographic fairness and prospective clinical evaluation should be investigated in future research. It is concluded that overall, NeuroFusion offers the groundwork for creating a more accurate, transparent and clinically relevant AI-powered assessment for Alzheimer's disease.

REFERENCE

[1] Viswan, V., Shaffi, N., Malathy, E., Chemmalar Selvi, G., Kavitha, B. R., Abdesselam, A., ... & Mahmud, M. (2026). Multimodal fusion and explainability of artificial intelligence models in Alzheimer's Disease detection. *Brain Informatics*, *13*(1), 5.

[2] Ahmad, I., Tarannum, T., & Uddin, M. M. (2026). Machine Learning and Deep Learning for Alzheimer's Disease Diagnosis: A Survey of MRI, PET, Multimodal Fusion, Transformers, Graph Neural Networks, and Explainable AI. *American Journal of Advanced Technology and Engineering Solutions*, *6*(01), 622-649.

[3] Shukla, G., Awasthi, V., Dubey, P., Nipane, S., Roy, S., Iyer, R., & Yogarayan, S. (2026). NeuroFusion for Robust and Explainable Multimodal Deep Learning in Fine-Grained Staging of Alzheimer's Disease Across Imaging and Clinical Biomarkers. *Engineering, Technology & Applied Science Research*, *16*(1), 32195-32202.

[4] Öztürk Söylemez, D., & Ay Doğru, S. (2026). NeuroFusion-ViT: A Hybrid CNN–EVA Transformer Model with Cross-Attention Fusion for MRI-Based Alzheimer's Stage Classification. *Diagnostics*, *16*(5), 754.

[5] Awais, M. (2026). *A Multimodal Feature Fusion and Explainable AI Driven Framework for Early Detection and Staging of Dementia* (Doctoral dissertation, CAPITAL UNIVERSITY).

[6] Ansari, A. S., Mohammadi, M. S., Cattani, C., & Tassaddiq, A. (2025). An advanced multimodal image fusion model for accurate detection of Alzheimer's disease using MRI and PET. *Frontiers in Medical Technology*, *7*, 1699821.

[7] Gasmi, K., Ammar, L. B., Krichen, M., & Alghuried, A. (2026). FLAME: Federated Learning and Aggregated Multi-Model Ensemble for Multi-Class Alzheimer's Disease Stage Classification from Structured Clinical Data. *Diagnostics*, *16*(13), 2029.

[8] Tiwari, S., Atluri, V., Kaushik, A., Yndart, A., & Nair, M. (2019). Alzheimer's disease: pathogenesis, diagnostics, and therapeutics. *International journal of nanomedicine*, 5541-5554.

[9] Tahami Monfared, A. A., Byrnes, M. J., White, L. A., & Zhang, Q. (2022). Alzheimer's disease: epidemiology and clinical progression. *Neurology and therapy*, *11*(2), 553-569.

[10] Breijyeh, Z., & Karaman, R. (2020). Comprehensive review on Alzheimer's disease: causes and treatment. *Molecules*, *25*(24), 5789.

[11] Khan, S., Barve, K. H., & Kumar, M. S. (2020). Recent advancements in pathogenesis, diagnostics and treatment of Alzheimer's disease. *Current neuropharmacology*, *18*(11), 1106-1125.

[12] Safiri, S., Ghaffari Jolfayi, A., Fazlollahi, A., Morsali, S., Sarkesh, A., Daei Sorkhabi, A., ... & Kolahi, A. A. (2024). Alzheimer's disease: a comprehensive review of epidemiology, risk factors, symptoms diagnosis, management, caregiving, advanced treatments and associated challenges. *Frontiers in Medicine*, *11*, 1474043.

[13] Asuku, A. O., Ayinla, M. T., Gbonjubola, O. O., Babatunde, S. S., Olajide, T. S., & Oyerinde, T. O. (2024). Neurocognitive aspects of dementia. In *Nutrition in brain aging and dementia* (pp. 109-130). Singapore: Springer Nature Singapore.

[14] Beata, B. K., Wojciech, J., Johannes, K., Piotr, L., & Barbara, M. (2023). Alzheimer's disease—Biochemical and psychological background for diagnosis and treatment. *International journal of molecular sciences*, *24*(2), 1059.

[15] Abubakar, M. B., Sanusi, K. O., Ugusman, A., Mohamed, W., Kamal, H., Ibrahim, N. H., ... & Kumar, J. (2022). Alzheimer's disease: an update and insights into pathophysiology. *Frontiers in aging neuroscience*, *14*, 742408.

[16] Anwal, L., Chandakavate, S., & Lalitha, S. (2021). A comprehensive review on Alzheimer's disease. *World J Pharm Pharm Sci*, *10*(7), 1170.

[17] Marcos-Rabal, P., Gonzalez-Fuentes, J., Castro-Vazquez, L., Lozano, M. V., Rodriguez-Robledo, V., Santander-Ortega, M. J., ... & del Mar Arroyo-Jiménez, M. (2021). Neurodegenerative diseases: a multidisciplinary approach. *Current pharmaceutical design*, *27*(30), 3305-3336.

[18] Afifi, I., Elgendy, M., Abdelfatah, M., & El-Sappagh, S. (2026). Vision and convolutional transformers for Alzheimer's disease diagnosis: a systematic review of architectures, multimodal fusion and critical gaps. *Brain Informatics*, *13*(1), 1.

[19] Ansari, A. S., Mohammadi, M. S., Cattani, C., & Tassaddiq, A. (2025). An advanced multimodal image fusion model for accurate detection of Alzheimer's disease using MRI and PET. *Frontiers in Medical Technology*, *7*, 1699821.

[20] Gayathri, T., Manjula, G., Kenchannavar, H. H., Sudha, D., Jankatti, S. K., Kaur, R., & Venkateswarlu, B. (2026). QuantumNeuroXAI: a quantum-inspired deep learning framework with explainability for brain signal analysis and neurological disorder detection. *Scientific Reports*.

[21] Sethuraman, R., & Padmakala, S. (2026, February). NeuroFusionNet: A Hybrid Deep Learning Framework for Multimodal Brain Tumor Classification. In *2026 International Conference on Visual Analytics and Data Visualization (ICVADV)* (pp. 1699-1706). IEEE.