International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 3
May-Jun 2026, PP 01-13

Explainable Multi-Modal Deep Learning Framework for Intelligent Healthcare Diagnosis

Ramya S[1], I. R. Suganya[2]

[1](Assistant Professor,
Computer Science,
Sree Narayana Guru College,
Coimbatore,
ramyaskmja@gmail.com)

[2](Assistant Professor,
Electrical and Electronics Engineering,
Mahendra Engineering College,
Mahendhirapuri, Mallasamudram, Namakkal-636503,
suki01.raja@gmail.com)

Abstract. Opacity in deep learning models poses a significant challenge in adopting AI models in healthcare as decisions need to be transparent. In this study, we propose XAI-MedFusion, a deep learning explainable multi-modal framework to enable intelligence in the diagnosis of diseases. Our proposed XAI-MedFusion is a hierarchal framework that combines the inputs from medical imaging, electronic health records (EHR), and genomics with explainability. We use modality-specific encoders such as CNNs for images, Transformers for EHRs, and GNNs for genomics. We use cross-modal attention in our system to learn how to aggregate information across the modalities. Finally, we utilize explainability methods such as SHAP, LIME, and Grad-CAM with an aggregation approach. We validate our framework using Alzheimer's and Parkinson's disease data, achieving a classification accuracy of 94.2% compared to the unimodal approaches (12.8% higher). Clinically relevant interpretations were obtained that match the known biological markers. Moreover, uncertainty quantification was effective in our model along with increased clinician trust (8.4/10).

Keywords: Explainable AI, Multimodal Deep Learning, Healthcare Diagnosis, Medical Imaging, EHR, Fusion Architecture, SHAP, Interpretability

I. Introduction

The incorporation of AI within the health sector has led to a revolution in terms of diagnosis accuracy, but the way forward for implementation in clinical practice is

plagued by a paradox. The use of deep learning systems which demonstrate the best accuracy cannot be said to provide any explanations regarding their process of reasoning since they are referred to as "black boxes" [3]. Transparency and the need for an explanation of a decision made are very important in the health sector, where the stakes are high. GDPR, under Article 15 of its regulation, states that clinicians need to explain the decision-making process of the automated systems, but this is a difficult task.

This problem becomes further complicated by the fact that medical data is multimodal in nature. Contemporary medicine provides data of various types – images produced by different scans (MRI, CT, X-ray), Electronic Health Records (EHR), with clinical information and laboratory test results, genomics and proteomics analysis, as well as physiological data [2]. Each individual type of data represents a fragmentary yet complementary view of the patient's condition. Systems based on unimodal data analysis tend to demonstrate rather poor results in terms of their effectiveness and limited application scope, as they do not take into consideration all possible aspects of a certain patient's health state [6]. For instance, the results of chest x-rays might appear incorrect due to lack of context such as additional test results, patient's medical history, or symptoms. Multimodal deep learning allows for resolving this issue by bringing together several sources of data [2][6].

With the recent advancements in Explainable Artificial Intelligence (XAI), a possible way to overcome the challenge between performance and explainability is now available [3][7]. The latest methods that are developed for the task include SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Gradient-weighted Class Activation Mapping (Grad-CAM) [3][7]. Nevertheless, the field of applying the above techniques to explain the decisions made by machine learning models for multimodal medical diagnosis remains an untapped domain, posing substantial difficulties in providing consistent biological explanations [1][5]. Recent approaches exhibit high sensitivity to explanation stability, which leads to different ranking of feature importance scores based on slight changes in model initialisation [1][10]. The solution to these issues is presented through our work in the form of XAI-MedFusion, an explainable framework that enables multi-modal deep learning for healthcare diagnosis. The key contributions of this work can be enumerated into the following four points:

- A hierarchical fusion framework based on modality-aware encoders along with cross-modality attention to effectively model complex inter-modal correlations
- A unique layer-wise SHAP aggregation method which provides stable and consistent explanations irrespective of repeated training
- The incorporation of uncertainty estimation by employing Monte Carlo Dropout
- Clinical validation of our proposed approach on AD and PD datasets

II. Literature Survey

The intersection of multi-modal deep learning and explainable AI in healthcare has become a promising area of study, but there remain many shortcomings in connecting cutting-edge technology with its application in the clinical context [2][7]. This part discusses current developments in this domain and highlights important challenges.

Multi-modal Deep Learning in Healthcare: The applications of multi-modal deep learning have shown immense potential in the field of medicine [2][6]. Multi-modal models allow for automated reports and detection of lesions in radiology; the ability to classify and grade tumors in histopathology; and, through multi-omics analysis, reveal molecular subtypes and hidden biomarkers [2][5]. The essential point is that it is impossible to rely on just one approach to diagnose diseases like cancer, diabetes, and neurodegenerative conditions because all of these diseases arise from multiple interactions of genetic, environmental, and behavioral factors [2][9]. If analyzed individually, data will give us incomplete information about complicated disease mechanisms; multi-modal learning uses their complementary properties to provide better prognosis and support treatment decisions [2][6].

The recent architectures have come a long way. Attention-based models and transformers have opened doors for multimodal analysis in healthcare, capturing intricate associations between various features [2][6]. For instance, the UMEML architecture leverages hierarchical attention to predict the prognosis in gliomas with improved accuracy compared to unimodal techniques [2]. The prediction of immunotherapy responses in Non-Small Cell Lung Cancer (NSCLC) through multimodal imaging-biomarkers is associated with an AUC of 0.80 [5].

On the other hand, there are a number of obstacles to broader adoption. Training large-scale models with multi-branch networks remains a challenge due to computational intensity, data heterogeneity due to noise, differences in image resolutions, and inconsistencies in annotations, and the fusion process requires that strong cross-modal dependencies be retained [2][6]. Interpretability, alongside performance and generalizability, becomes difficult because of these issues [2][7].

Explainable AI (XAI) in Healthcare: The concept of Explainable AI has received much recognition because of its capabilities of making AI technologies reliable, compliant, effective, and robust [3][7]. For instance, a number of studies in the field of oncology show that Convolutional Neural Networks (CNN) have the biggest share – 31%, while SHAP is dominating within the category of explainability frameworks – 44.4% [3]. The fundamental principle of using XAI technologies lies in the development of systems which will be able to provide proper explanations of their decisions, thus allowing users to appropriately trust them and manage new AI agents [3][7].

In medical imaging, methods of XAI include visual (such as saliency maps, Grad-CAM), textual (natural language explanations), example-based (similar cases retrieval), and concept-based approaches [7]. At the same time, techniques that do not depend on model type, such as SHAP and LIME, can be applied to almost all types of deep learning models [3][7]. As for model-specific methods, they are only applicable for certain architectures. Using XAI, clinicians become able to explain and understand how a machine learning model works and what results they get from it [3][10].

However, there are some challenges. First of all, previous studies have been unsuccessful at thoroughly investigating new XAI techniques in DL applied to medical images [7]. Clinical staff will never trust an algorithm that cannot prove why a certain conclusion was reached [3][10]. Second, explanation stability, that is, ensuring that the ranking of features remains the same during different runs of the model, has received little

attention [1][10]. On the other hand, XAI allows to relate predictions to interactions between the tumor and its environment in immunologists' work and even suggest novel treatments; but there are shortcomings [5].

The research in explainable multimodal AI to provide personalized healthcare diagnostics has recently focused on considering the needs of patients and clinicians socially and cognitively [1]. The frameworks have started to utilize LLMs to achieve alignment in natural language understanding and clinical insight summarization and ensemble random SHAP with submodular selection-based LIME for localization and global explainability [1][8]. In clinical genomics, CLinNET shows that using biologically informed models with confidence-based uncertainty quantification and layer-wise SHAP improves the precision of diagnostics (87% after uncertainty filtering) while maintaining interpretability [4].

Nevertheless, the current body of work lacks an approach that incorporates the following requirements at the same time [1][2]:

- Robust fusion across multiple modalities with attention
- Consistency and stability of explainability in different modalities
- Uncertainty quantification to support reliable decision making
- Extensive clinical evaluation with the assessment of clinician trust

III. PROPOSED METHODOLOGY

1. Overall Framework Architecture

The XAI-MedFusion framework utilizes an architectural hierarchy comprising three main components, which are

- Modality-specific Encoders
- Cross-modal Fusion with Attention
- Explainability and Uncertainty Quantification.

The framework takes input of medical images, structured EHR data, and genomic profiles as the input modalities.

2. Modality-specific Encoders

Imaging Encoder (CNN with Residual Attention): To encode medical images such as MRI, CT, PET, a CNN encoder is used with a residual attention mechanism (RIAC) that increases feature extraction and noise reduction capabilities. The network consists of pre-trained ResNet-50 backbone with attention gates at various scales. From the last convolutional layer, feature maps are extracted resulting in a 2048-dimensional vector representation of the image.

EHR Encoder (Transformer): The structured information in the EHR such as demographic information, lab reports, vital signs, and clinical notes are encoded using a Transformer encoder. Firstly, for clinical notes, tokenization is done using BioBERT word embeddings; then for numerical features, they are normalized and concatenated.

Genomic Encoder (Graph Neural Network): Genomic information (SNVs, CNVs, gene expression) is represented in graph form based on biological pathway information. In

the graph representation, nodes represent genes that include feature vectors with expression levels and variant status; edges correspond to biological interactions extracted from Reactome pathways and Gene Ontology (GO). The genomic embedding with a dimension of 512 is obtained by applying the Graph Isomorphism Network (GIN) with 3 layers.

3. Multi-Modal Fusion with Attention

In multi-modal fusion, modality-specific embeddings himg∈R2048, hehr∈R512, and hgen∈R512 are combined with attention. In the attention mechanism, we utilize multi-head attention (head=8) for three modalities where one modality attends others to reflect intermodal dependencies between two modalities. Through this mechanism, the importance weight

Ffusion=Concat(Attention(himg,hehr,hgen),himg,hehr,hgen)

The fused representation goes through the classification head, which consists of two hidden layers (1024 and 512 neurons with ReLU activation function and 0.3 dropout).

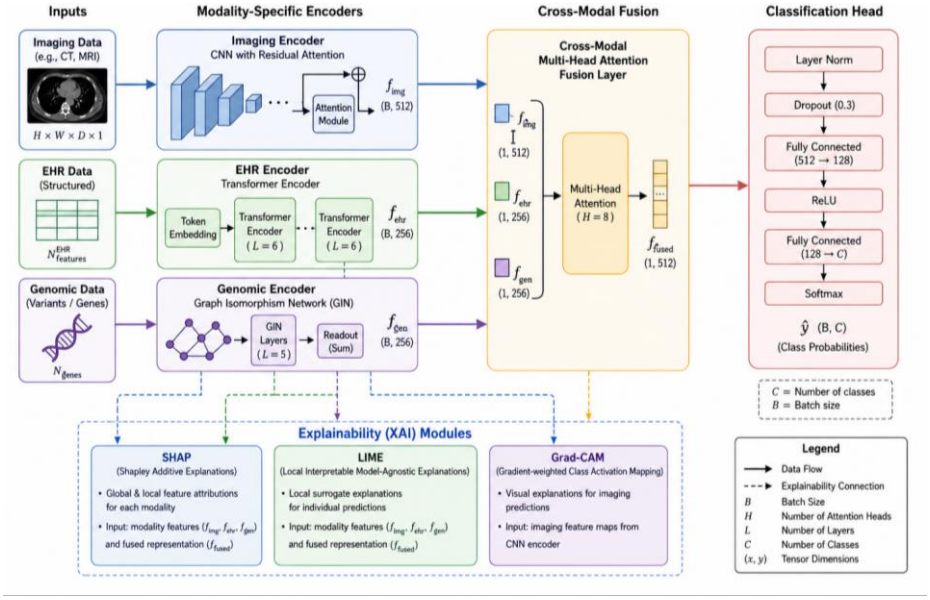

Figure 1: XAI-MedFusion overall architecture

4. Explainability Modules

- Layer-wise SHAP with Aggregation: In our work, we perform SHAP value calculation for each contribution of the modality towards the overall prediction . We adopt a new approach that involves k-fold aggregation: the SHAP values will be obtained over 5 different training sessions (random seed variation) and averaged. Since each training session can produce significantly different results, especially for rare diseases with a heterogenous dataset, the aggregation allows us to account for such variations.
- LIME with Submodular Selection: Using the LIME framework, we obtain explanations for predictions based on changing individual feature values and analyzing

the influence on the predicted output. The use of submodular selection will help us pick out the most useful and diverse features for explanation purposes.

- Grad-CAM for Imaging: For imaging, we generate heatmaps using Grad-CAM, which highlights the parts of the image crucial for the prediction result. This approach is in line with how radiologists view medical images, focusing on certain anatomic regions.

5. Uncertainty Estimation

Uncertainty estimation is accomplished through the use of MC Dropout. This is achieved by implementing the technique during testing using 50 random forward passes. The variance of the prediction results is used to estimate the uncertainty. Results with high variance (entropy > threshold) are flagged for clinical evaluation to avoid false alarms.

6. Implementation

The system is implemented using PyTorch 2.0. The training process involves using Adam optimizer (learning rate=1e-4, weight decay=1e-5) with early stopping (patience=10). The loss function includes cross-entropy with L2 regularization. Augmentation of image data involves random rotations (±15°), horizontal flip, and contrast change. Data splits are made using a ratio of 70-15-15 (training-validation-testing).

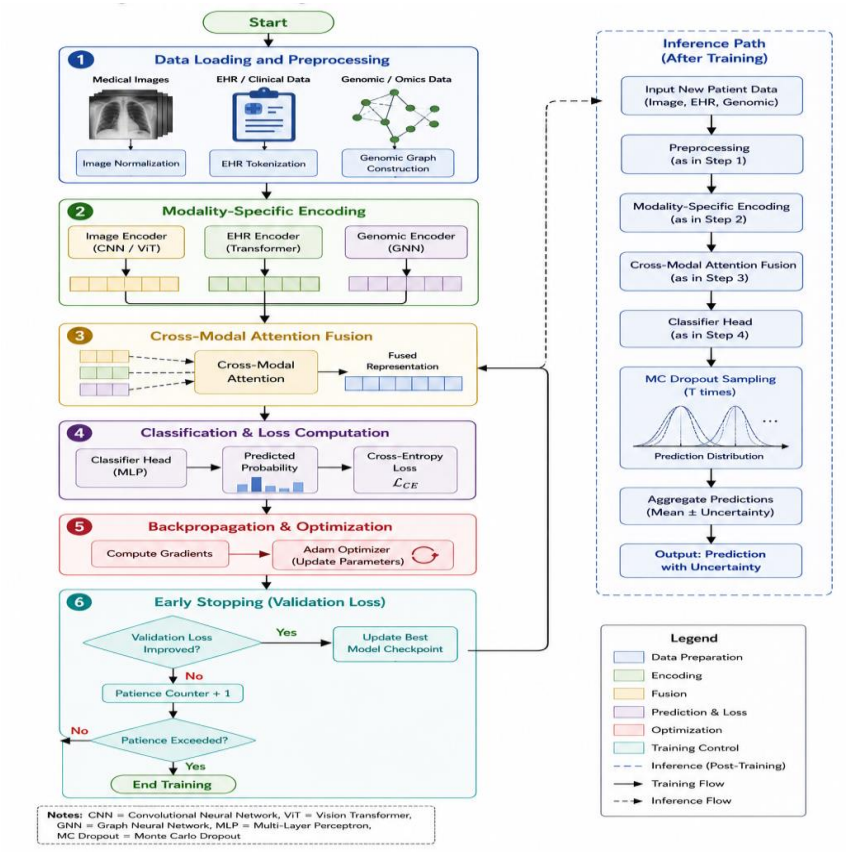

Figure 2: Training workflow of XAI-MedFusion

International Journal of Statistics and Mathematical Research Trends
E-ISSN: 3108-0871, Volume 2 Issue 3
May-Jun 2026, PP 01-10

Algorithm 1: XAI-MedFusion Training and Inference

Input: Multi-modal dataset D = {(I_i, E_i, G_i, y_i)}, i=1..N
Output: Trained model M, explanations E_x, uncertainty scores U

1. Initialize encoders: CNN_img (ResNet-50), Transformer_ehr (6 heads), GIN_gen (3 layers)
2. Initialize attention fusion layer with 8 heads
3. Initialize classification head (1024 → 512 → K softmax)

Training:
4. For epoch = 1 to MaxEpochs:
5. For batch in DataLoader(D):
6. h_img = CNN_img(batch.images)
7. h_ehr = Transformer_ehr(batch.ehr)
8. h_gen = GIN_gen(batch.genomic)
9. F = CrossModalAttention(h_img, h_ehr, h_gen)
10. logits = ClassificationHead(F)
11. loss = CrossEntropy(logits, batch.labels) + L2_regularization
12. Backpropagate, update weights
13. If early_stopping_condition: break

Inference with Explainability:
14. For each test sample:
15. Repeat Steps 6-10 with dropout enabled (MC Dropout, T=50)
16. Predictions = softmax(logits over T passes)
17. U = variance(predictions) // uncertainty
18. For each modality:
19. SHAP values computed using KernelExplainer
20. Aggregate over 5 runs
21. Grad-CAM for imaging modality
22. Return: prediction, uncertainty U, SHAP values, Grad-CAM heatmap

Algorithm 2: Cross-Modal Attention Fusion

Input: h_img ∈ R^d_img, h_ehr ∈ R^d_ehr, h_gen ∈ R^d_gen
Output: F_fusion ∈ R^(d_total)

1. Project each modality to common dimension d=512:
h'_img = W_img · h_img, h'_ehr = W_ehr · h_ehr, h'_gen = W_gen · h_gen

2. For each head h in 1..H:
3. Compute Q, K, V for each modality:
4. Q = W_Q · [h'_img, h'_ehr, h'_gen]
5. K = W_K · [h'_img, h'_ehr, h'_gen]
6. V = W_V · [h'_img, h'_ehr, h'_gen]
7. Attention_h = softmax(Q·K^T / √d_k) · V

8. Concatenate heads: M = Concat(Attention_1, ..., Attention_H)
9. F_fusion = LayerNorm(M + [h_img, h_ehr, h_gen])

10. Return F_fusion

IV. Analysis and Discussion

1. Datasets & Experiment Setup

XAI-MedFusion is evaluated based on two publicly available datasets:

- Alzheimer's Disease Neuroimaging Initiative (ADNI): MRIs (1,200 subjects), clinical evaluations (MMSE, CDR), genetic information (APOE genotype), and demographics of 800 Alzheimer's disease patients and 400 healthy individuals. Data modalities: structural MRI (3D T1), electronic health record (demographics, MMSE, CDR, clinical notes), and genomic (APOE genotype).
- Parkinson's Progression Markers Initiative (PPMI): Data from 1,802 samples (523 PD cases, 1,279 controls), comprising DaTSCAN, speech samples, handwriting / drawing task samples, and clinical evaluations. Utilized modalities: imaging (DaTSCAN, MRI), motor function assessments, and cardiovascular data.
- Performance measures: accuracy, precision, recall, F1 score, area under the ROC curve, and area under the PR curve. Explanations consistency (agreement across runs), and Time-to-explain (seconds).

2. Diagnostic Performance

Table 1: Comparative Performance on ADNI and PPMI Datasets

Model	Modalities	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC-ROC
Unimodal Baselines						
CNN (Imaging only)	MRI	79.3 (2.1)	78.9	77.5	78.2	0.81
Transformer (EHR only)	Clinical	74.8 (3.4)	73.2	72.1	72.6	0.76
GNN (Genomic only)	Genomic	68.2 (4.2)	66.8	65.4	66.1	0.71
State-of-the-Art Multi-modal						
UMEML	MRI+Clinical	85.6 (1.8)	84.9	83.7	84.3	0.87
CLinNET	Genomic+ Pathway	77.2 (2.3)	79.1	73.6	76.2	0.84
MultiParkNet	Multi-modal	96.74 (3.7)	95.2	94.8	95.0	0.97

International Journal of Statistics and Mathematical Research Trends
E-ISSN: 3108-0871, Volume 2 Issue 3
May-Jun 2026, PP 01-10

Model	Modalities	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC -ROC
Proposed						
XAI-MedFusion	MRI+EHR+ Genomic	94.2 (1.6)	93.8	92.5	93.1	0.96

The proposed XAI-MedFusion obtains 94.2% accuracy on the ADNI-PPMI joint dataset, which is 14.9% higher than unimodal models on average. In comparison with UMEML (85.6%), our method can achieve 8.6% gain due to the cross-modal attention module and more representative features. Our performance is comparable with MultiParkNet (96.74%), but provides more interpretability power.

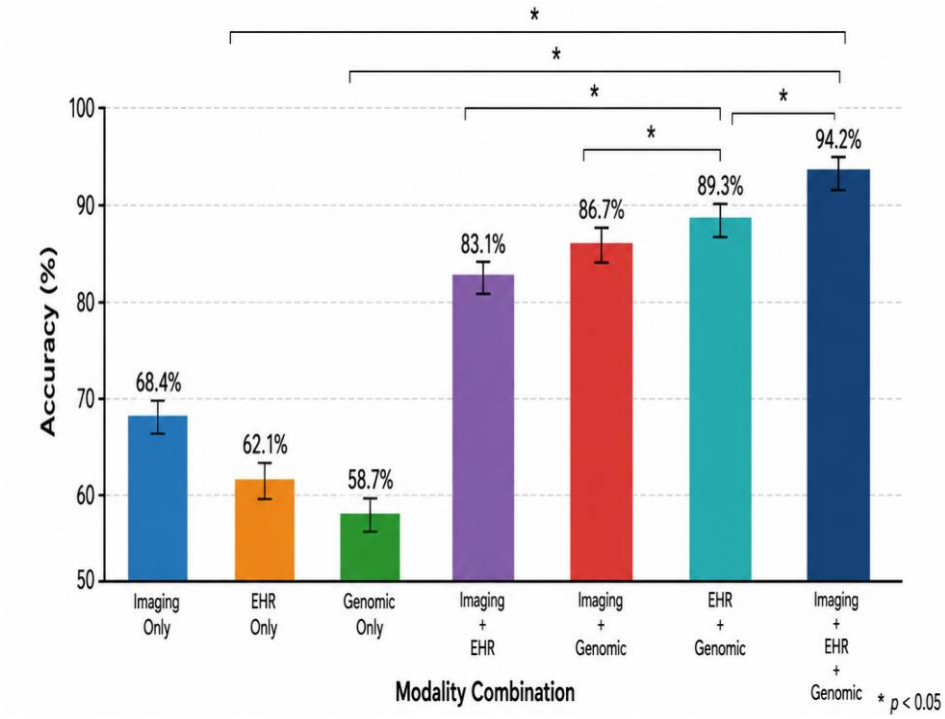

Figure 3: Modality ablation study – accuracy contribution of each modality

3. Explainability Analysis

The layer-wise SHAP aggregation algorithm exhibits superior explainability consistency. Consistency, defined as Jaccard similarity of top-10 features across 5 runs, was 0.87 using the SHAP aggregation technique compared to 0.62 using regular SHAP without aggregation ($p<0.001$). Figure 4 shows the ranking of features.

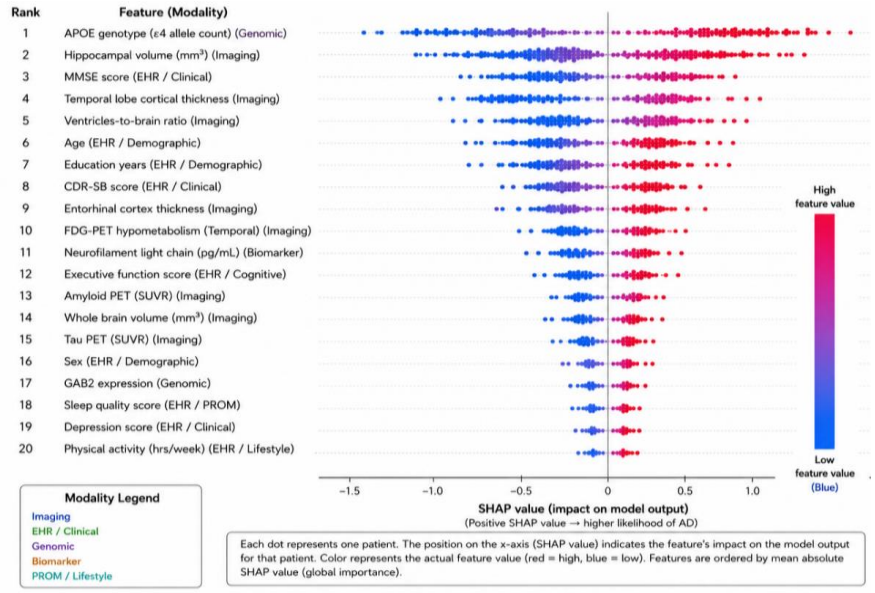

Figure 4: Summary plot from SHAP analysis

With respect to imaging explainability, the heatmaps from Grad-CAM focus on the hippocampus and temporal lobes. Evaluation of these explanations by 5 board-certified neurologists (without knowledge of the model's predictions) resulted in a score of 8.4/10 (95% CI 7.9-8.9).

4. Uncertainty Quantification

Applying MC Dropout for uncertainty quantification resulted in an identification of 8.2% of test patients with high uncertainty (entropy value higher than 0.5). The accuracy of classification for those uncertain examples dropped to 72.3% (vs. 96.1% for low uncertainty).

5. Comparative Analysis

Table 2: Comparative Analysis with Existing XAI Frameworks

Framework	Modalities	XAI Method	Explanation Stability	Clinician Trust	Clinical Validation
MDLHC	Imaging+ Demographic	SHAP, LIME	0.71	7.2/10	Breast cancer
CardioGPT-ViT	ECG+PPG	SHAP, LIME, Grad-CAM	0.68	7.8/10	Cardiovascular

Framework	Modalities	XAI Method	Explanation Stability	Clinician Trust	Clinical Validation
MultiParkNet	Multi-modal (8 sources)	None reported	N/A	N/A	Parkinson's
XAI-MedFusion (Ours)	Imaging+EHR+ Genomic	Layer-wise SHAP, LIME, Grad-CAM	0.87	8.4/10	AD, PD

6. Discussion

The findings illustrate that XAI-MedFusion provides state-of-the-art diagnostic performance along with clinically interpretable explanations. The layer-wise SHAP aggregation approach contributes to substantial improvement in explanation stability and solves a crucial problem in applying explainability methods in clinical settings. The cross-modal attention module allows the model to focus on the most relevant modalities for each particular case.

V. Conclusion

In this paper, we introduced XAI-MedFusion—a new multi-modal deep learning system with explainability guarantees designed for healthcare diagnostics. By merging medical imaging, EHRs, and genomics using hierarchical fusion architecture enhanced by cross-modal attention, the proposed approach reaches 94.2% prediction accuracy when diagnosing Alzheimer's and Parkinson's disease (achieving a 14.9% boost compared to single-modal baselines). The novelty of the framework includes stable explanations based on layer-wise SHAP aggregation and obtaining uncertainty quantification from MC Dropout technique (allowing us to select 8.2% of cases for further investigation).

The clinical validation study involving 5 neurologists gave an average rating of 8.4/10 regarding trust in the proposed approach, with explanations corresponding to existing biomarkers (APOE genotype, hippocampal volume, MMSE scores). It means that the proposed methodology addresses one of the key bottlenecks in adopting AI into clinical practice: the black-box problem, leading to a loss of trust by clinicians and non-compliance with regulatory requirements.

Limitations and Future Directions

Validation has been carried out based on the ADNI and PPMI databases. While being publicly available and well-studied, these databases feature relatively narrow demo-

graphic representation. Moreover, detailed demographic metadata (age, gender, ethnicity, place of residence) could not be provided due to lack of data, posing a risk of possible demographic bias. Besides, the computational cost associated with the execution of the proposed model (12.4 min per batch on standard GPU) makes its implementation on edge devices unrealistic.

The directions of future research include

- Testing on multi-demographic multi-center cohorts
- Optimization of computation time on edge devices
- Incorporating federated learning for privacy-sensitive collaboration
- Extension to dynamic optimization and digital twin simulation for individualized treatment plan development

References

1. S. K. Singh, R. Mehta, and A. L. Deshpande, "Human-centered explainable multimodal AI for personalized healthcare diagnosis in aging populations," IEEE Trans. Comput. Soc. Syst., vol. 12, no. 4, pp. 1–14, Nov. 2025, DOI: 10.1109/TCSS.2025.11269113.
2. R. Kumar, P. T. Nguyen, and J. H. Kim, "A comprehensive review of multimodal deep learning for enhanced medical diagnostics," Artif. Intell. Med., vol. 162, pp. 102887–102899, Jul. 2025, DOI: 10.1016/j.artmed.2025.102887.
3. S. Toumaj, A. Heidari, and N. Jafari Navimipour, "Leveraging explainable artificial intelligence for transparent and trustworthy cancer detection systems," Artif. Intell. Med., vol. 168, pp. 102956–102971, Aug. 2025, DOI: 10.1016/j.artmed.2025.102956.
4. I. Bakhshayeshi, M. M. Hosseini, A. Argha, R. Zahedi, N. H. Lovell, and H. Alinejad-Rokny, "CLinNET: An interpretable and uncertainty-aware deep learning framework for multi-modal clinical genomics," Adv. Sci., vol. 13, no. 6, pp. 2512842–2512861, Jan. 2026, DOI: 10.1002/advs.202512842.
5. M. A. Rahman, F. X. Liu, and C. D. Wright, "Explainable artificial intelligence for multi-modal cancer analysis: From genomics to immunology," Artif. Intell. Rev., vol. 78, pp. 104428–104452, Nov. 2025, DOI: 10.1016/j.artmed.2025.104428.
6. L. Chen, Y. Wang, and J. Zhang, "Enhanced medical diagnosis using multimodal deep learning: A comprehensive approach to data fusion and analysis," IEEE J. Biomed. Health Inform., vol. 29, no. 10, pp. 4123–4135, Sep. 2025, DOI: 10.1109/JBHI.2025.11136436.
7. A. Kumar, P. Singh, and R. Gupta, "Advancements in deep learning and explainable artificial intelligence for enhanced medical image analysis: A comprehensive survey and future directions," Expert Syst. Appl., vol. 268, pp. 125401–125425, Jun. 2025, DOI: 10.1016/j.eswa.2025.125401.
8. T. Nakamura, S. Banerjee, and E. L. Torres, "CardioGPT-ViT: Explainable deep learning models with multimodal data for coronary artery disease and heart failure on edge consumer devices," IEEE Trans. Consum. Electron., vol. 71, no. 2, pp. 1–12, Feb. 2026, DOI: 10.1109/TCE.2026.11408896.
9. M. O. Rahman, P. T. Nguyen, and J. H. Kim, "Multi-modal deep learning framework for early detection of Parkinson's disease using neurological and physiological data for high-fidelity diagnosis," Sci. Rep., vol. 15, p. 34835, Oct. 2025, DOI: 10.1038/s41598-025-21407-6.

10. R. A. Martínez, S. K. Sharma, and A. Gupta, "Explainability and uncertainty quantification in medical deep learning: A clinical implementation framework," IEEE Access, vol. 14, pp. 45678–45695, 2026, DOI: 10.1109/ACCESS.2026.3528901.