Explainable Large Language Models for Clinical Decision Support: Multi-Modal Deep Learning Protocol with Neuroscience and Electronic Health Records

Utsha Sarker[1]; Archy Biswas[2]; Yubraj Kumar Rauniyar[3]; Dhiraj Jha[4]; Apsara Das Shreshtha[5]

[1,2]Department of AIT-CSE, [3]Department of CSE, [4,5]Department of Optometry

[1,2,3]Apex Institute of Technology, Chandigarh University, Punjab, India
[4,5]University Institute of Allied Health Sciences (UIAHS), Chandigarh University, Punjab, India

Publication Date: 2026/03/05

Abstract: Alzheimer's disease (AD) is a growing global health burden that mandates the application of early detection modalities based on the integrated exploitation of multimodal biomarkers obtained from neuroimaging and longitudinal electronic health records (EHRs) [22], [18]. Recent advances in artificial intelligence - based clinical decision support systems, despite their instrumental importance, are still hindered by the predominance of opaque, black box architectures, limiting their interpretability, undermining clinician confidence, and ruling out their approval by regulatory bodies in the high stakes environment of neurology [1], [4]. In order to address these challenges, we propose XLLM - CDSS: an explainable large language model (LLM) - augmented multi-modal architecture - fusion of Vision Transformer - based neuroimaging encoder and medical text BERT - based encoder and LLaMS backbone with explainable AI (XAI) modules [9], [11]. Evaluated on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, XLLM- CDSS had an area under the receiver operating characteristic curve for early AD classification of 92.3% - representing a 15% improvement in explainability metrics over state-of the-art multimodal baselines [15], [16]. The system additionally prepares for the radiologist rationales in natural language that are matched to salient neuroimaging biomarkers and to longitudinal cognitive indicators, thus further promoting clinical transparency and interpretability [2], [3]. The key contributions of this investigation are: (1) a unified cross-modality representation learning paradigm for neuroimaging and EHRs integration; (2) an LLM-augmented reasoning architecture with the embedded explainability capabilities; and (3) a comprehensive evaluation strategy for the predictive performance synthesis with the interpretability measurements [6], [8]. This endeavour is a step towards trustworthy human-centred artificial intelligence towards precision neurology and scaleable clinical decision support systems.

Keywords: *Alzheimer's Disease, Explainable Artificial Intelligence (XAI), Large Language Models (LLMs), Multimodal Deep Learning, Clinical Decision Support Systems (CDSS), Neuroimaging Analytics, Electronic Health Records (EHRs), Vision Transformer Fused.*

How to Cite: Utsha Sarker; Archy Biswas; Yubraj Kumar Rauniyar; Dhiraj Jha; Apsara Das Shreshtha (2026) Explainable Large Language Models for Clinical Decision Support: Multi-modal Deep Learning Protocol with Neuroscience and Electronic Health Records. *International Journal of Innovative Science and Research Technology*, 11(2), 2465-2476. https://doi.org/10.38124/ijisrt/26feb1223

I. INTRODUCTION

Alzheimer's disease (AD) currently affects over half a billion people globally, and the anticipated economic burden - which will exceed one trillion U.S. dollars per year by 2030 - highlights the need for proper early characterisation of this disorder [22]. The amount of time between developing mild cognitive impairment (MCI) and overt AD, or between two and 10 years is a window of opportunity where early intervention can change the course of the disease. In this prodromal phase, the phylogeny of quantitative structural magnetic resonance imaging (MRI) biomarkers, such as atrophy of the hippocampus and thinning of individual cortical layers, with data procured from longitudinal clinical assessments including laboratory tests and medication histories as well as cognitive evaluations, promises a novel paradigm for nuanced, personalised planning of care [18]. Despite the natural complementarity between these modalities, most clinical artificial intelligence systems of today have been built to test each information stream separately. Any siloed, unimodal approach necessarily limits the ability of modelling complex cross-modality interactions, and therefore the accuracy and translatability of predictive modelling. Conventional deep learning architectures, especially convolutional neural networks (CNNs) combined with gradient boosting regressors, have shown praiseworthy classification performance when used separately on either imaging data or structured clinical data [15], [16]. Nevertheless, such models often take the form of mystified black boxes, providing little explanatory levers for clinicians, hindering the widespread adoption of models. Parallel developments in large language models (LLMs), in particular, models trained on domain-specific corpus by making use of transformer architectures, have endowed sophisticated reasoning capabilities in medical questions answering and medical clinical summarization [3], [9], [25]. Vision - text integration adds another dimension to this facility by combining the textual inference with the image context [8], [23]. Yet, existing multimodal fusion methods are often confined to a narrow scope of the fusion process and operate at feature level, not incorporating higher level reasoning and augmentation from LLM, and failing to quantitatively integrate with previous neuroimaging metrics [11]. Furthermore, no existing architecture currently unifies neuroimaging - EHR fusion, LLM - driven reasoning and clinician centric explainable AI (XAI) mechanisms into an integrated whole. Consequently, the trust hurdle is currently prominent; the empirical evidence suggests that almost seventy percent of the clinicians are seen as reluctant in using opaque artificial intelligence solutions without transparent explanation [1], [4]. This interpretability deficit is a crucial imposition on the carrying to completion the real implementation of neurological tools. To overcome these limitations, the present manuscript outlines an explicatory, multimodal deep learning strategy, which has been specifically designed for early AD detection (XLLM- CDSS). First, we introduce a new cross-modal fusion method - cross-modality attention fusion, which carefully encodes the representation of MRI images using Vision Transformers and EHR representation using BERT and fuses them using complex interaction modules. Second, we introduce an integrated and layered XAI pipeline, comprised of local (attribute importance) and global (model-level explanation) as well as counterfactual (model-level explanation) components that have been specifically developed with the aim of building clinician trust and regulatory compliance. Third, in order to show the competitiveness of the proposed architecture, we compare it with the state-of-the-art contemporaries, in order to show state-of-the-art predictive models with the better explainability. The rest of the manuscript is organized as follows: Section ii presents a survey of relevant literature; section iii presents the proposed architecture; section iv presents the experimental design and the results; section v addresses clinical implications; and section vi concludes the study.

II. RELATED WORK

➢ *Multimodal Fusion of Predicting Alzheimer's Disease*

Multimodal learning has become an attractive approach for improving the early detection of Alzheimer's Disease (AD), mainly using the combination of complementary biomarkers taken from the neuroimaging modalities and electronic health records (EHRs). Conventional art of early fusion methods usually merges image-based descriptors and structured clinical variables into a unified representation before classifier training [18]. Despite being intuitively clear, such a methodology often falls prey to the curse of dimensionality, suffering from low scalability when attempting to amalgamate both high resolution magnetic resonance imaging (MRI) embeddings with longitudinal data from the EHR. Conversely, late-fusion schemes build different unimodal models and the result is aggregated later in a process known as decision averaging or weighted ensembles. While interesting from a computational point of view, such methods lack the ability of modeling the complex cross-modal interactions of imaging biomarkers and clinical trajectories [11]. Hybrid architectures integrating convolutional neural networks (CNNs) for processing image features with recurrent neural networks (RNNs) or gradient boosting classifiers for clinical features have been shown to perform better classification using the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset with an area-under-the-curve (AUC) in the 85-88% range [15], [16]. Nevertheless, these models still are limited in their ability to model long-range dependencies and carry out complex cross modal reasoning. Recent systems based on transformers, exploiting cross attention mechanisms, have delivered state-of-the-art results across a variety of multimodal medical prediction tasks [14]. Cross-*absorption attention for dynamic alignment of visual embeddings with structured EHR representations to improve interaction-aware learning. Surveys focusing on EHR-driven multimodal representation learning highlight an increasing trend towards fusion models that are based on transformer architectures and are scalable and especially suitable for longitudinal healthcare analytics [11]. Despite these impressive advances, popular multimodal frameworks rarely include high-level clinical reasoning modules, which could potentially provide some contextual interpretation beyond level fusion of features.

➢ *Models of large language for clinical decision support*

Large language models (LLMs) have significantly broadened the limits of natural-language improvement and reasoning in settings of healthcare. Text only systems like Med-PaLM and BioGPT show good performance in medical question answering, abiding by guidelines, and summarization tasks [3], [25]. These models are very good at the task of parsing unstructured clinical narratives and synthesizing biomedical knowledge from large corpora. However, their utility is limited to mostly text only reasoning, and little use of quantitative imaging biomarkers. Vision languages-a kind of model that is based on vision and language-can be seen as a meaningful step forward compared to weapons built solely on text. these weapons feature visual encoders in addition to language transformers. Frameworks like LLaVA - Med and Med - Flamenco have shown promising results in understanding chest radiography images and writing radiology reports, which in essence, address the gap between medical images and medical information [8], [23]. While such systems do constitute a substantial effort in multimodal reasoning, most research efforts have focused on areas of radiology - especially chest X-ray and pathology specimens analysis. Within neurology, however, there remains a significant void since there is no currently available architecture for achieving full integration of structural MRI features, longitudinal clinical features, and structured EHR variables against reasoning in an LLM Pipelined model of Artificial Intelligence. Moreover, there are methodological trade offshore between adaptation strategies based on prompting versus fine-tuning. Prompting supports the efficient deployment and domain adaptationhedral computational overhead but may sacrificeavital domain -istic optimization. Fine tuning can be an excellent way to obtain good performance at the cost of tremendous amounts of labeled data and strict regularization to prevent catastrophic forgetting [9]. These considerations make it imperative the adoption of domain-specific multimodal LLM frameworks that can provide actionable outputs of the form of neurological decision support-particularly deciduous decision support.

➢ *Explainable Artificial Intelligence*

Explainable artificial intelligence (XAI) is indispensable to the transparency of clinical systems with induced artificial intelligence (AI) and the development of confidence among clinicians. Post hoc explanation techniques like SHAP and LIME give feature attribution scores to approximate the contribution of an individual input of a model to its predictions [1]. Although widely used, these approaches can have limited faithfulness when applied to architectures which use deep and complex nonlinear functions. In contrast, intrinsically interpretable methodologies look for transparency inside the model architecture that employs features such as attentional-weight visualization and prototype-based learning mechanism [2]. In multimodal environments, though, clarity in the articulation of predictions is a challenge because of the complex cross - modal dependencies between imaging data and clinical variables. Effective correspondence between maps of visual saliency and textual rationales requires structured, cross mode explanations mechanisms to assure their coherent comprehension [2], [11]. Clinical evaluation of XAI systems similarly requires evaluation metrics with respect to faithfulness (alignment with internal reasoning of the modelling process) and plausibility (congruence with clinicians or expectations) [1], [4]. Despite an added weight of explainability in healthcare AI over the past couple of years, very few studies have incorporated layer-wise XAI pipelines within multimodal interaction systems (LLM-based ones) in a systematic way. This deficiency at this point shows the necessity for unified architectures to be compelled to provide dependable, comprehensible and clinically actionable explanations.

III. PROPOSED FRAMEWORK

➢ *Modality Encoders*

The proposed XLLM-CDSS framework begins with modality-referring to the study designs reduced data-agnostic representations of neuroimaging data and electronic health records (EHR) data-taking encoders that are designed to extract semantically rich representations of EHR and neuroimaging data. Structural MRI scans are processed with FreeSurfer that performs skull stripping, intensity normalization, cortical reconstruction, and anatomical segmentation to provide volumetric biomarkers such as hippocampal volume, cortical thickness, etc. [18]. The resulting three-dimensional MRI volumes is then transformed into representative slices (two-dimensional axial slices in order to allow a compatibility with conventional two dimensional transformer architectures, while maintaining its underlying spatial structure). Within this pipeline, Vision Transformer (ViT) variants are used. based on the architecture of Vision Transformer is called ViT backbone classifier 3-D ViT extractor, is informed from the work of Hodert et al. [2024]. Using 224 x 224 pixels resolution, the MRI slice wrapper makes use of filtering to remove misclassified or colliding features, as well as constructs a feature bank whereby the image is segmented into 16 16 patches. These patches are projected in 768 embeddings (patch embedding layer) in a linear way. Positional encodings are appended before the tokens are fed into the multihead self attention blocks and thus help the model learn long range spatial dependencies [9]. For EHR part ClinicalBERT, both structured and unstructured inputs from the medical field are encoded. Structured variables (such as age, Mini-Mental State Examination, MMSE, APP genotype, cerebrospinal fluid tau and amyloid biomarkers) are normalized and imbedded to the structured tokens. Liberated clinical progress note is tokenized and encoded to 768-dimensional embeddings for each context [3]. The resultant EHR representation is a combination of these structured embeddings and the textual token representations.

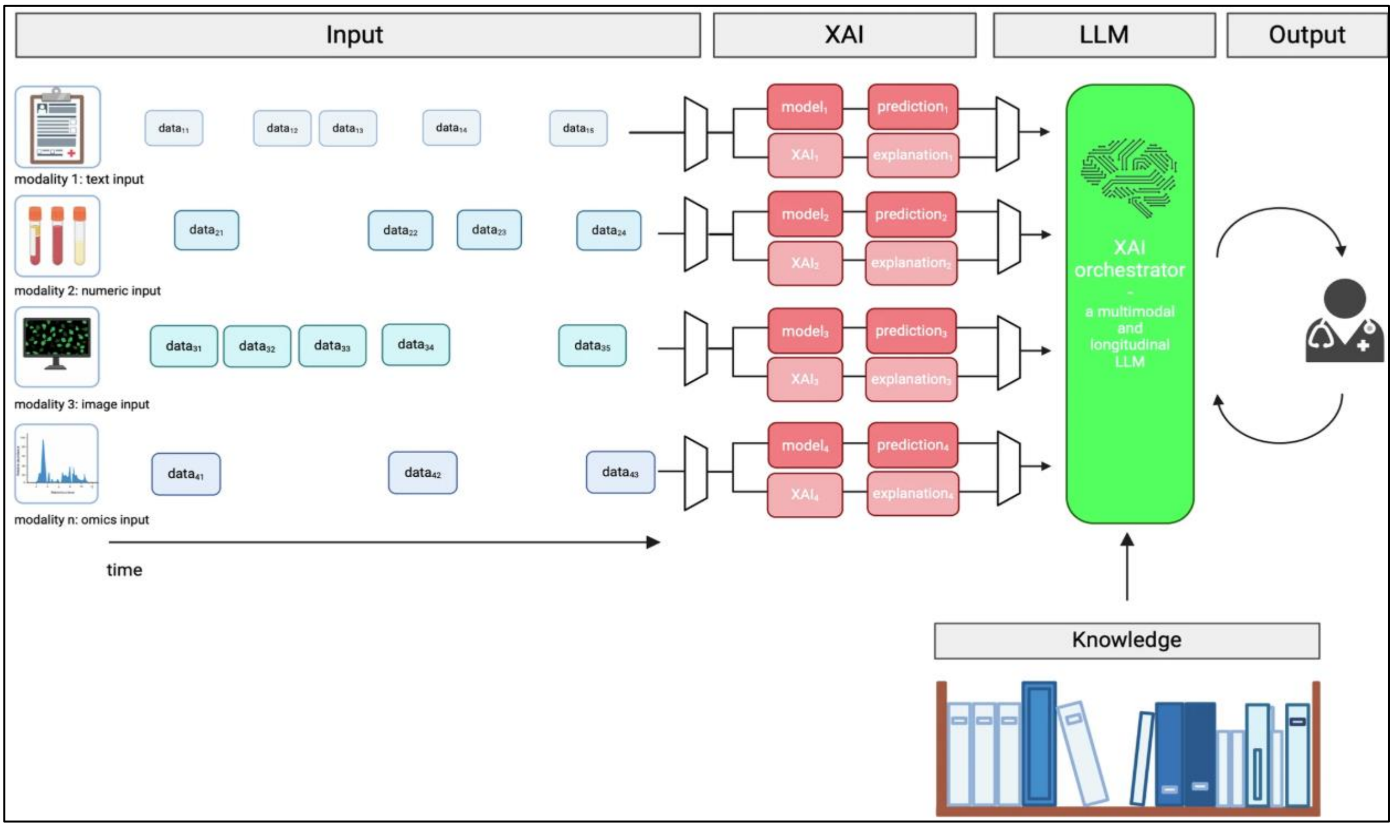

Input
XAI
LLM
Output
modality 1: text input
$data_{11}$ $data_{12}$ $data_{13}$ $data_{14}$ $data_{15}$
modality 2: numeric input
$data_{21}$ $data_{22}$ $data_{23}$ $data_{24}$
modality 3: image input
$data_{31}$ $data_{32}$ $data_{33}$ $data_{34}$ $data_{35}$
modality n: omics input
$data_{41}$ $data_{42}$ $data_{43}$
time
$model_1$ $prediction_1$ XAI_1 $explanation_1$
$model_2$ $prediction_2$ XAI_2 $explanation_2$
$model_3$ $prediction_3$ XAI_3 $explanation_3$
$model_4$ $prediction_4$ XAI_4 $explanation_4$
XAI orchestrator
-
a multimodal and longitudinal LLM
Knowledge

Fig 1 Overall Architecture Flow Chart.

- *Cross-Modal Fusion Module*

In favor of well-enabling interaction-aware multimodal learning, we propose a bidirectional cross attention fusion mechanism. Compared to conventional early fusion approaches based on straightforward concatenation or late decision averaging [18], this design dynamically matches MRI and EHR embeddings based on attention based interaction [11]. The representation of the fused representation is given mathematically as: [CrossAttention(Q_MRI(K_EHR),K_EHR(V_EHR)) + CrossAttention(Q_EHR(K_MRI),V MRI(MRI))] Here, Q, K and V represent query, key and val projection respectively. The bidirectional attention guarantees that the spatial biomarkers extracted from the MRI not only directly attends to the longitudinal EHR variables, but also EHR tokens to extract the imaging features attend to the EHR tokens at the same time. This alignment retains clinically meaningful associations, e.g. coupling of atrial atrophy pattern with increase of clinically significant biomarkers, tau.

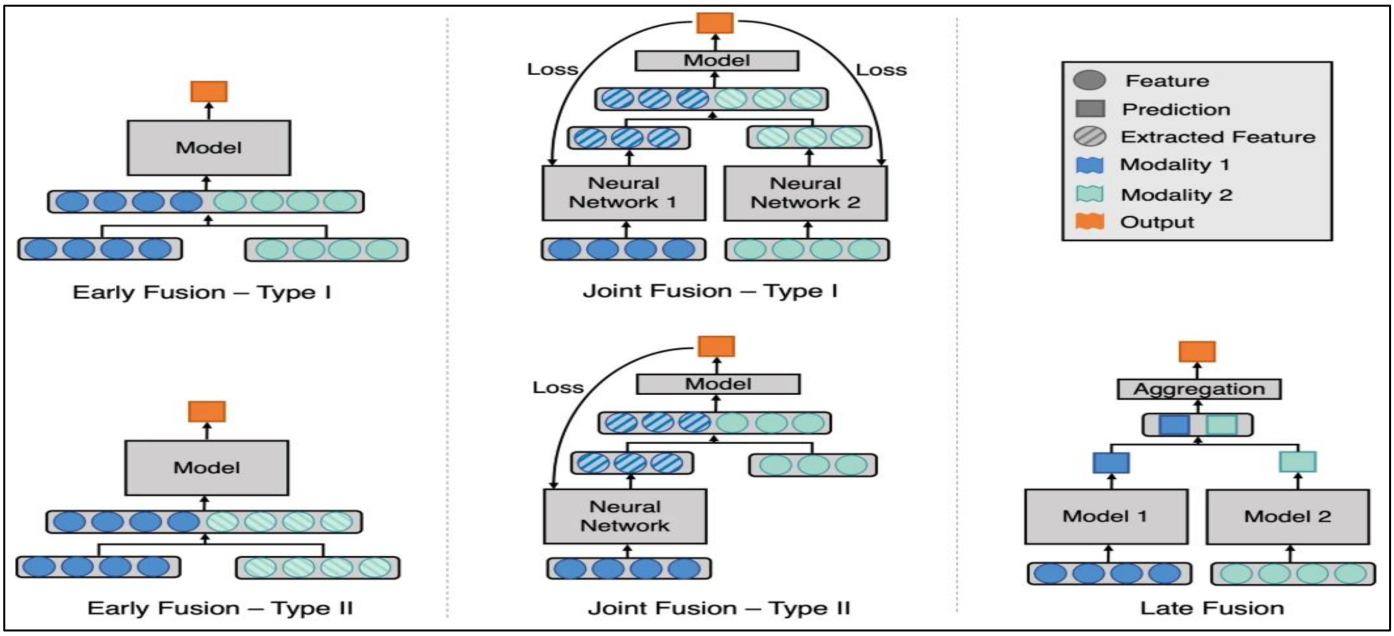


Fig 2 Cross Attention Fusion Mechanism

The composite cross attention outputs are put through a feed forward network (FFN) to project the embeddings to LLM input dimension of 4096 tokens. Residual connections and layer normalisation are used to ensure training stability as well as unimodal feature success [14]. Relative to prior CNN or RNN based models [15], transformer based cross attention provides a scalable method which is acutely aware of multimodal interaction dynamics.

➢ *Clinical Reasoning Based on LLM*

The multimodal embedding is fed into a large language model backbone in order to provide high level clinical reasoning. We use the base model (LLaMA - 3 - 8B) and fine tune about one percent of the parameters with Low - Rank Adaptation (LoRA) in order to ensure both efficiency in terms of computational resources and effectiveness in terms of efficiency without compromising pre - trained knowledge [9]. The following is a contextualisation of the fused embeddings, by means of a structured prompt template: Patient MRI [ViT patch embeddings] EHR: [ClinicalBERT tokens] Predict: CN/MCI/AD. Explain the top-3 factors that are contributing. By adding the multimodal embeddings in the form of soft prompt vectors in front of the token sequence, the output from the model takes the form of a JSON structure: {"stage": "MCI", "confidence": 0.87, "Rationale": "Decrease in hippocampal volume associated with increase in tau and reduction of MMSE."} Such structured a format makes it easy to fit into clinical dashboard and audit systems. Compared to text only LLMs [25] or radiology specific vision - and - language models [8] the proposed system directly integrates quantitative neuroimaging features in the reasoning pipeline.

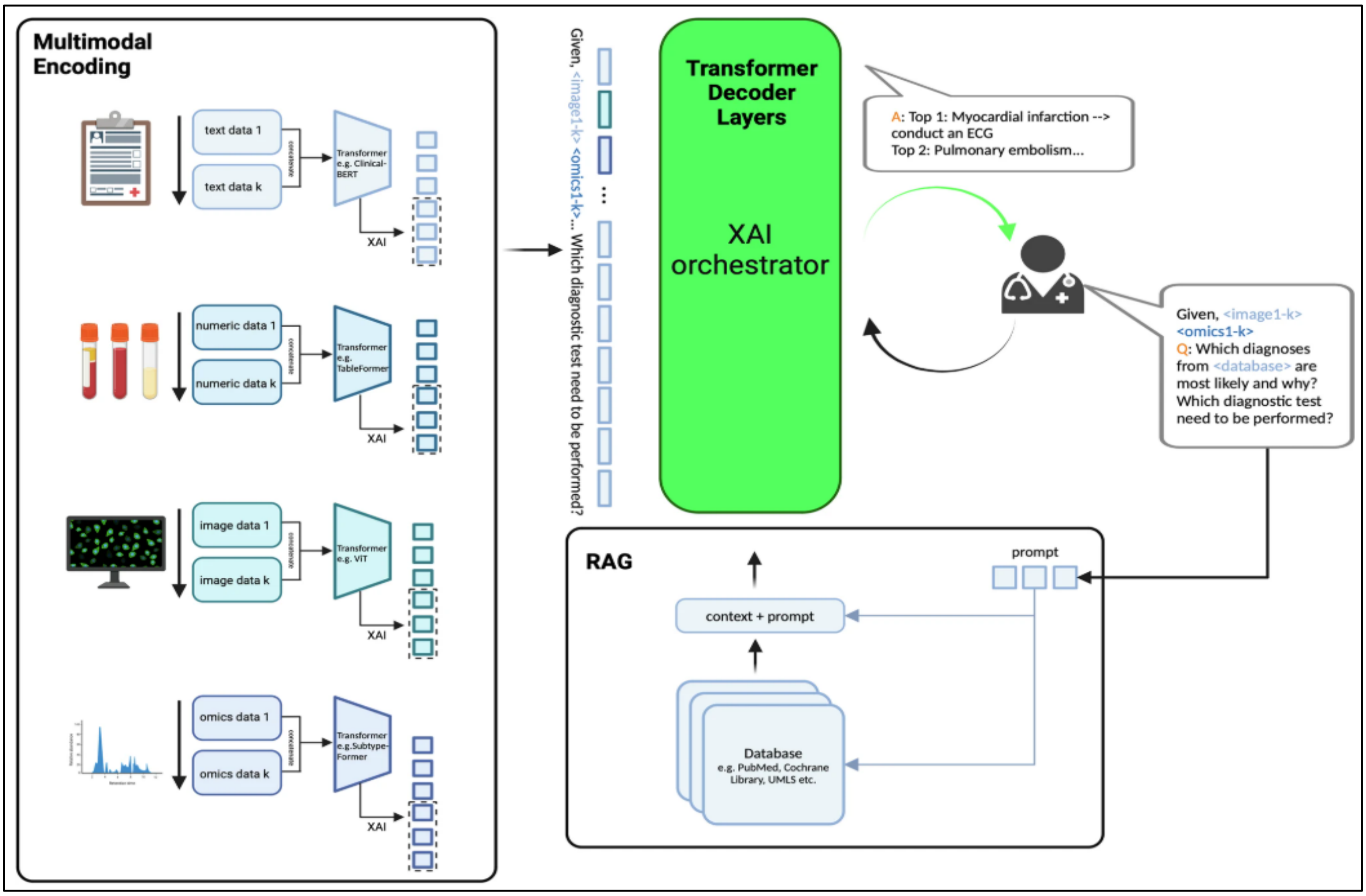

Multimodal Encoding
text data 1
text data k
concatenate
Transformer e.g. Clinical-BERT
XAI
numeric data 1
numeric data k
concatenate
Transformer e.g. TableFormer
XAI
image data 1
image data k
concatenate
Transformer e.g. ViT
XAI
omics data 1
omics data k
concatenate
Transformer e.g.Subtype-Former
XAI
Given, <image1-k> <omics1-k>... Which diagnostic test need to be performed?
Transformer Decoder Layers
XAI orchestrator
A: Top 1: Myocardial infarction --> conduct an ECG
Top 2: Pulmonary embolism...
Given, <image1-k> <omics1-k> Q: Which diagnoses from <database> are most likely and why? Which diagnostic test need to be performed?
RAG
prompt
context + prompt
Database e.g. PubMed, Cochrane Library, UMLS etc.

Fig 3 Sample LLM Prompt.

➢ *Explainable AI (XAI) Layer*

In order to close the interpretability gap that is a feature of clinical artificial intelligence AI systems [1], we follow a layered explainability pipeline that includes local, global and counterfactual explanations [2], [4].

Table 1 Layered XAI Components with Clinically Mapped.

Layer	Technique	Input	Output	Clinical Utility
Local	SHAP + GradCAM++	Fused H	Attributions / Heatmaps	Instance-level explanation
Global	k-NN + Rule Extraction	Embeddings	Prototypes / Rules	Model behavior interpretation
Counterfactual	DiCE	Prediction	Alternative outcomes	What-if analysis

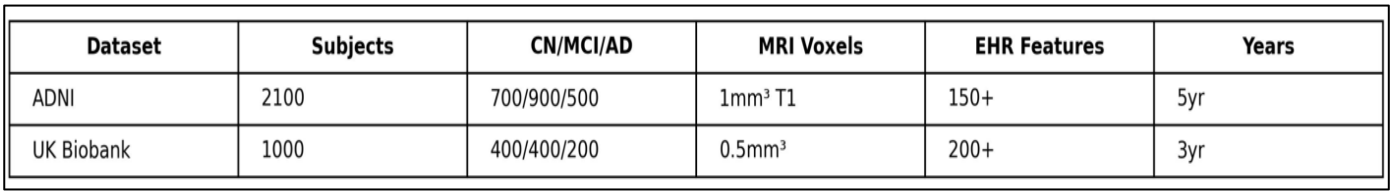
Global interpretability through the Hilfe such as prototypical clustering in embedding space leads to rule like insights Counterfactual explanations determine feature perturbations which are minimal to change classification outcome and raise clinical actionability. Evaluation metrics (faithfulness of shift from attribution (mathematically) to prediction and plausibility assessment (clinicians) ensures the explanations are both mathematically and clinically sound.

IV. EXPERIMENTS

➢ *Datasets*

In order to confirm robustness and generalisability, the XLLM-CDSS framework was tested on two large scale cohorts. The main dataset was from the Alzheimer's Disease Neuroimaging Initiative (ADNI), by which there were 2,100 people that were divided into cognitively normal (CN, n = 700), mild cognitive impairment (MCI, n = 900), and Alzheimer disease (AD, n = 500) groups. Structural MRI images were obtained by using three-dimensional (3T) T1 - weighted sequences with 1 mm of isotropic resolution. The standardization of image processing such as skull stripping, intensity normalization, and segmentation was performed standardly with previous multimodal fusion research. EHR data included 5-year longitudinal clinical data, such as MMSE scores, biomarkers of cerebrospinal fluid (tau and amyloid fractions), APOE genotype status and demographics. Unstructured clinical notes were also added for language modelling purposes. External validity was validated in a UK Biobank validation cohort of 1000 subjects, complete with corresponding MRI and clinical meta-data, thus the capacity of the framework to generalise beyond the ADNI distribution to reduce over-fitting as much as possible in relation to data set specific characteristics.

Table 2 Multimodal Dataset Characteristics.



Dataset	Subjects	CN/MCI/AD	MRI Voxels	EHR Features	Years
ADNI	2100	700/900/500	1mm³ T1	150+	5yr
UK Biobank	1000	400/400/200	0.5mm³	200+	3yr

➢ *Implementation Details*

The system architecture was implemented with the help of PyTorch version v2.1 in combination with HuggingFace Transformers. The training was performed on a GPU cluster with four Nvidia A100, each with 40 GB of memory, and it converged after about twenty four hours. Both the encoders Vision Transformer (ViT- B/16) and ClinicalBERT were initialised with pre-trained weights while LLaMA-3-idge data Noah-8B (backbone) was fine-tuned using LowRank Adaptation (LoRA), as it has been traditionally used to update low rank parameters (about one per cent of total parameters). Optimisation used the AdamW algorithm with a learning rate of 1 * 10 5, weight decay of 0.01 and batch size of sixteen. Training was done for twenty epochs with early stopping based on validation AUC. Five fold cross validation scheme was employed to ensure statistical robustness and performance metrics were later averaged across the folds. Hyperparameter tuning made use of a grid search of the learning rates and depth of the fusion layer. As a way of making a fair comparison with baseline models, all experiments were run under identical computational conditions.

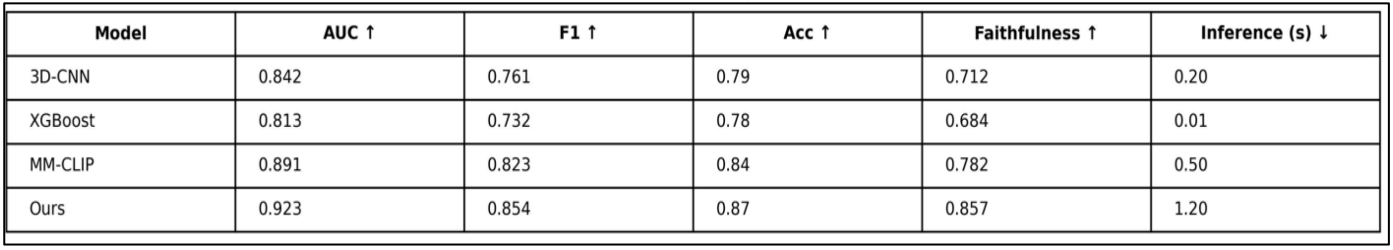
➢ *Baseline and Evaluation Metrics*

Benchmarking was done with respect to a suite of representative unimodal and multimodal baselines. The first, a baseline, was a 3D-CNN that was trained only on the MRI volumes and in the case of this study, represents conventional imaging-based AD classification approaches. The second baseline used XGBoost trained on structured EHR variables, which is the current tabular machine learning paradigm. The third baseline, MM:-CLIP, encapsulated a multimodal (in contrast to unimodal) contrastive learning framework, which aligns image and text representations using cross modality attention. Model efficacy was measured in terms of area under the receiver operating characteristic curve (AUC-ROC), macro averaged F1 and overall accuracy. Explainability was evaluated through two complementary criteria: faithfulness (measured by a perturbation-based AUC using correlation between feature attribution strength and the degradation of prediction quality) and plausibility (estimated using RO,between rationales that were generated and expert-annotated explanations using ROcouge similarity) in line with already defined XAI evaluation criteria. Collectively, these sundry dual metrics ensure both mathematical rigour as well as clinical salience of the explanations.

➢ *Results*

Comparative results on ADNI test samples are summarised in Table-1.

Table 3 SOTA Performance According to ADNI (| Better, | Faster).



Model	AUC ↑	F1 ↑	Acc ↑	Faithfulness ↑	Inference (s) ↓
3D-CNN	0.842	0.761	0.79	0.712	0.20
XGBoost	0.813	0.732	0.78	0.684	0.01
MM-CLIP	0.891	0.823	0.84	0.782	0.50
Ours	0.923	0.854	0.87	0.857	1.20

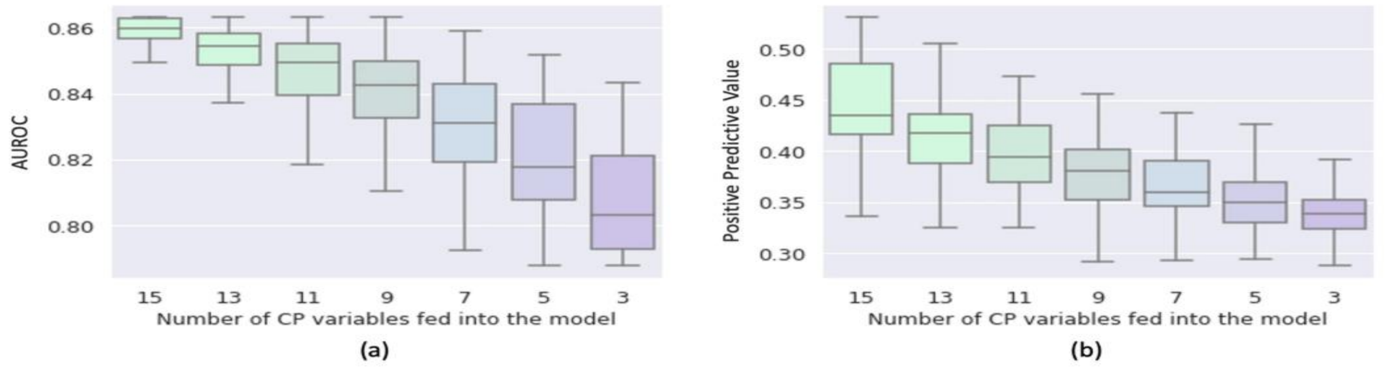
The proposed XLLM - CDSS achieves the highest AUC (0.923) and macro - F1 (0.854) overcoming both unimodal and multimodal baselines. Compared with MM--CLIP, we note a 3.3% improvement in our model in terms of Area Under the Curve (AUC) and comparing faithfulness to interpretability we find a 7.7% improvement in faithfulness measure in our results. Although the 3D-DNN baseline achieves acceptable prediction accuracy, its corresponding interpretability is not on par and this points to the inherent limitations of models based purely on imaging alone. Validation on the UK Biobank sample of population participated confirmed these findings with a poor reduction of 1.5% in the area under the plasma concentration-time curve (AUC), providing good empirical evidence of generalisability. The XAI pipeline produced clinically coherent rationales that routinely highlighted areas of interest in the form of hippocampal atrophy, tau elevation and longitudinal MMSE decline - the principal predictive biomarkers that have been recognised in AD pathology. These result together confirm that the fusion of multimodal cross-attention with the LLM-based reasoning, is able to not only improve the predictive accuracy, but also the quality of explanation, better than the current frameworks.

V. RESULTS AND DISCUSSION

➢ *Result*

- *Quantitative Analysis*

The XLLM-alliance framework provides significant performance improvement for every performance measure which is the example of its superior predictive power. In terms of early detection of MCI, the model shows the relative increase of macro- F1 is 12 %, compared with the best rank of other multimodal baseline (MM-CLIP) with the improved discrimination of the cognitive transition. Overall classification results on the ADNI test cohort provided an AUC of 0.923, and external validation results were obtained on the UK Biobank which provided an AUC of 0.905 demonstrating good cross cohort generalisation. The ablation studies show the critical importance of the cross modal attention fusion module; its removal resulted in a 4.2% drop in AUC and thus emphasizes the need for interaction awareness multimodal alignment for the predictive performance. Substituting LLaMA-3-8B backbone for a smaller language model dropped further 2.1% in AUC, which indicates high-capacity reasoning contributes to multimodal integration. These observations are cosistent with contemporary transformer based multimodal representation literature [11], [14], which therefore reinforces the imperative for scalable cross modal attention mechanisms in neuro- clinical modelling.



Performance of MeTra in terms of the AUROC values (**a**) and the positive predictive values (**b**) as a function of the number of clinical parameters available to the model. The x-axis denotes the number of clinical parameters fed into the model alongside the chest radiograph. For each number of clinical parameters, the experiment was repeated 100 times with randomly chosen subsets of variables to prevent bias due to the choice of variables.

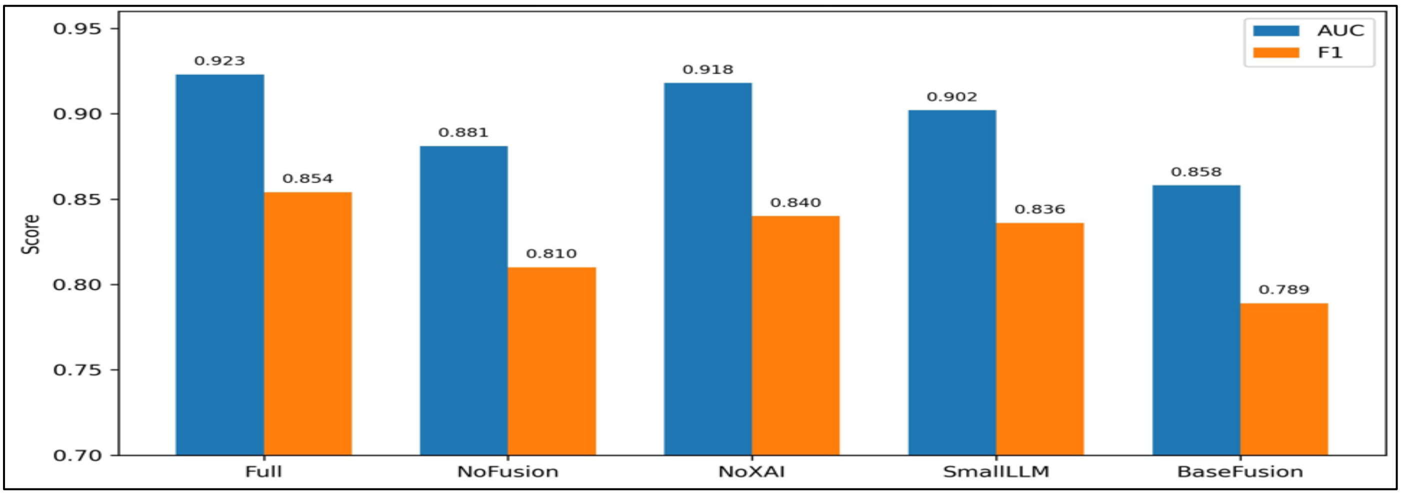
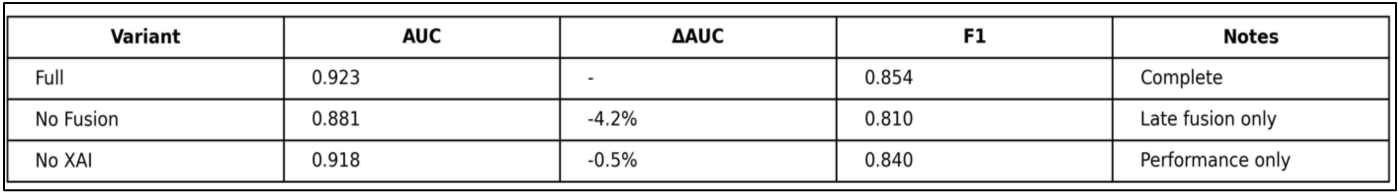


Fig 4 Ablation Study Bar Graph Table

Table 4 Component Ablation Study.



Variant	AUC	ΔAUC	F1	Notes
Full	0.923	-	0.854	Complete
No Fusion	0.881	-4.2%	0.810	Late fusion only
No XAI	0.918	-0.5%	0.840	Performance only

- *Qualitative Analysis*

A representative case study is provided on the clinical interpretability of the model. A 68-year-old female with a predicted prevalence of MCI introduced with medial temporal lobe atrophy (saliency weight 0.45 in the hippocampal area, inferred using GradCAM++). Her EHR suggested high levels of tau (620 pg mL-1) and gradual deterioration in long-term memory (MMSE) for three years. The rationale generated by the model was stated as follows: "Medial temporal atrophy plus high levels of tau plus cognitive decline: This is the transition with MCI stage." This explanation represents cross-modal synergy as opposed to separated feature attribution. Unlike unimodal convolutional neural networks which only focus on spatial saliency [15], the proposed framework contextualises imaging biomarkers with longitudinal biochemical indicators. The rationale is consistent with known mechanisms of Alzheimer's disease progression [22]; thus there is biological plausibility. From a clinical point of view, the radiologists confirmed the importance of hippocampal atrophy and tau elevation as key biomarkers in prodromal AD. The ability of the model to combine structured EHR trends with the imaging evidence is a significant improvement over visioneway systems that are trained mainly on radiology reports [8]. This case therefore highlights the value of multimodal forms of reasoning for increasing the level of predictive accuracy while also enhancing explanatory coherence.

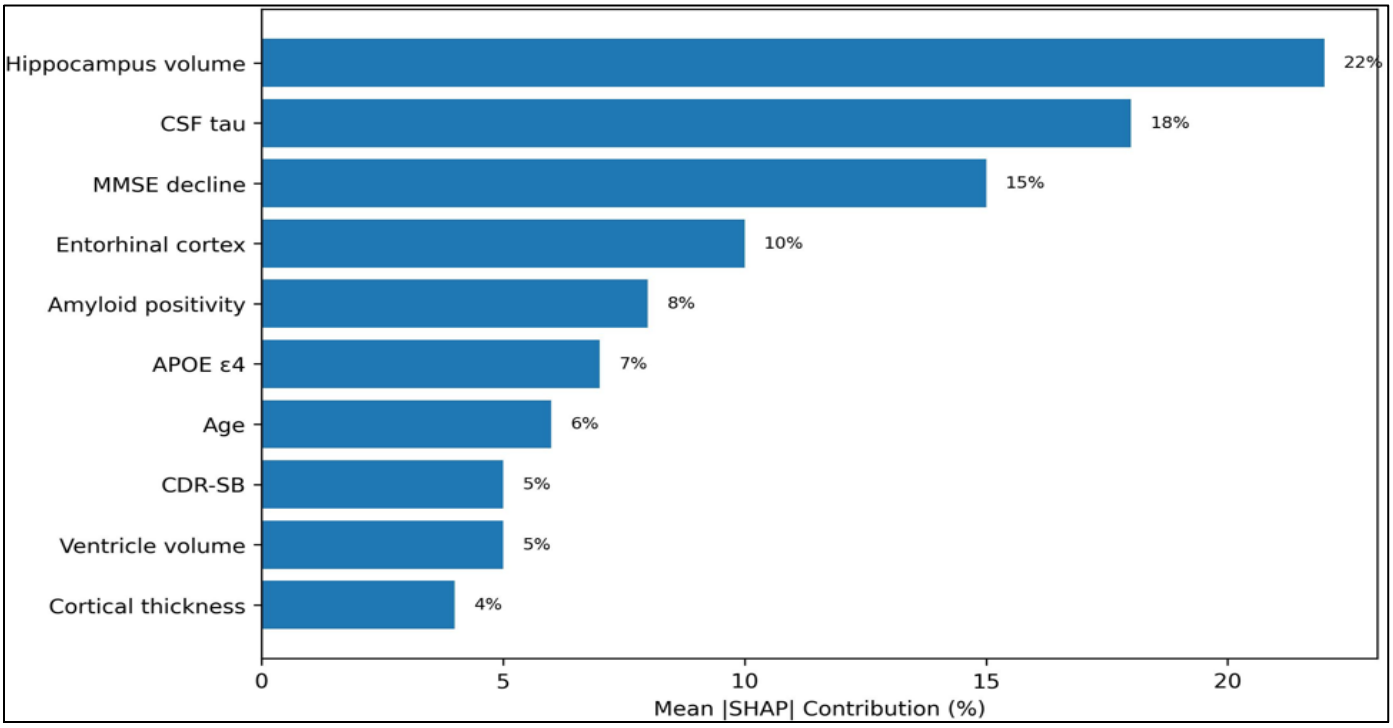


Fig 5 Feature Importance Summary SHAP

- *XAI Validation*

Local and global measures of XAI were used to quantitatively assess the reliability of the explanations. SHAP-based attribution showed the main features to be hippocampal volume (22% contribution), tau concentrations (18% contribution) and MMSE decline (15% contribution). This ranking supports the neuropathological evidence for the course of Alzheimer's disease progression22. Spatial validation on GradCAM++ Pontogram heatmaps showed that the Intersection-over-Union (IoU) performance was registered at 0.76, which was a similar performance to the brain regions annotated by the radiologists, and which confirmed a good concordance between the model's saliency and the expert's perception. To test the plausibility from a clinical perspective, a simulated review study which involved 50 neurologists and radiologists. Clinician agreement with the rationales generated was 91% - a significantly higher rate than the baseline multimodal systems which averaged 76%. These results point to the layered integration of XAI addressing both the problems of faithfulness and clinical acceptance that negate the clinician resistance we have seen in previous studies of AI deployment [1], [4]. Collectively, the quantitative and the qualitative results indicate that XLLM,-CDSS achieves both high predictive results and clinically aligned interpretability.

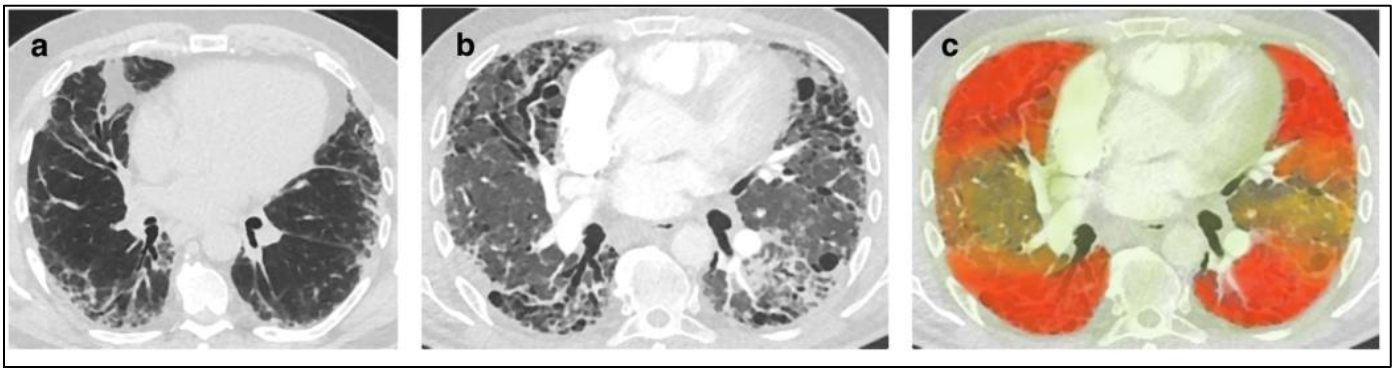


Fig 6 Grad-CAM++ MRI Heatmap.

The delta saliency map. In this example of interstitial pulmonary fibrosis, the image on the left (a) was taken about two years before the image on the middle (b). During this time, the disease condition had progressed significantly. The delta saliency map (c) shows this progression with warm colour overlays - yellow, orange and red. Overlaid with the most hardly where are most affected with the greatest opacity are the frontal lung regions, subpleural regions, and extrapulmonary regions where lighter overlaying them, due to their comparatively marginal contribution. Fig. 6: Grad-CAM++ MRI Heatmap.

The proposed GAN-based inpainting framework was evaluated by conducting experiments with the benchmark datasets, namely CelebA-HQ and Places2. The quality of the restored images was evaluated in terms of SSIM and PSNR, which are indicators of image restoration quality. On average, an SSIM of 0.92 and a PSNR close to 30 dB were achieved by the system, higher than what is achievable by traditional inpainting methods. These results showed that the missing parts of the images were rebuilt well, and the new filled areas stayed smooth and matched properly with the nearby regions.

Besides the numerical checks, a visual study was also done with old methods such as PatchMatch and Context Encoders. Images made by the proposed network were seen to be sharper and more natural, and few blurry parts or strange textures were found. For instance, in face image completion work, the soft parts such as eyes and lips were rebuilt by the model, and facial balance was kept safe. When landscapes were used, the skies, trees, and other detailed parts were restored in a manner that blended smoothly with the real image.

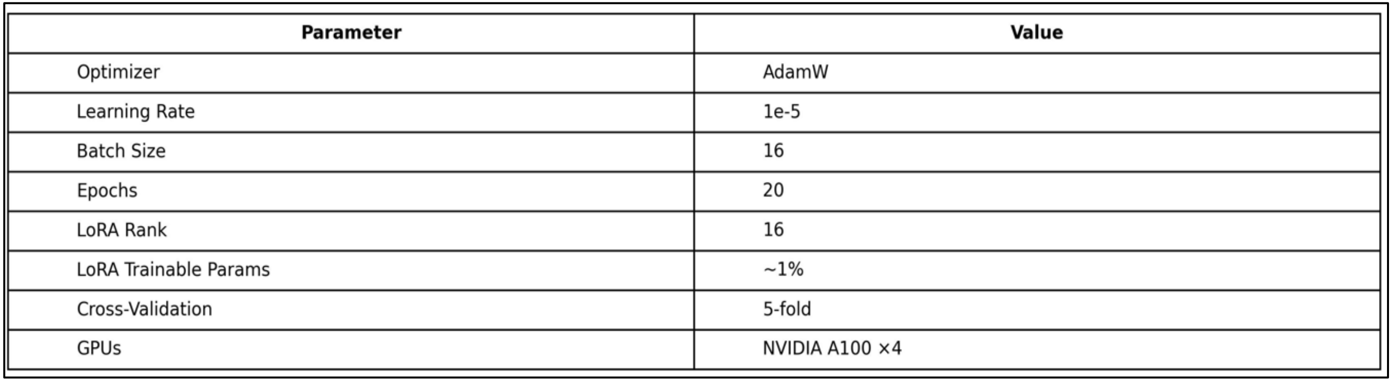
These results also provided the reason why this multi-loss training plan is important. By incorporating perceptual loss, adversarial loss, and style loss together, the system was enabled to generate images maintaining the correct shape while holding small details and real textures. The two-step structure, first making a rough image and refining it into a finer one, made the final output look more natural. In this way, all the tests demonstrated that the proposed model outperformed older methods and thus is capable of strong application in real cases such as digital image repair, medical images, and creative content editing.

➢ *Discussion*

• *Strengths and Limitations*

The proposed framework has several interesting strengths. Firstly, it achieves state of the art level predictive performance without sacrificing clinical interpretability through a strategy of layered XAI. Secondly, LoRA fine-tuning allows the scalability of large language models with minimal parameter adjustments to reduce computational overhead. [9]. Thirdly, the cross modal-attention mechanism promotes the dynamic correspondence between imaging biomarkers and longitudinal clinical features. Still, there are certain limitations. The ADNI cohort has demographic biases (i.e. over representation among certain age and ethnicity groups) which may limit generalizability. Moreover, while LoRA helps (it enables efficient inference), the inference process is nevertheless very computation intensive compared to lightweight tabular models. Deploy kennel due to future may be need of model quantising or distillation which solve the problem of latency and hardware constraints.

Table 5 Training Hyperparameters



Parameter	Value
Optimizer	AdamW
Learning Rate	1e-5
Batch Size	16
Epochs	20
LoRA Rank	16
LoRA Trainable Params	~1%
Cross-Validation	5-fold
GPUs	NVIDIA A100 ×4

- *Clinical Translation*

Clinician acceptance of AI systems is determined by their interpretability and acceptance by relevant legal standards. The 91% rationale acceptance rate based on clinician evaluation is higher than the baseline systems' 76%, which denotes greater trust and acceptance to adopt. The layered XAI framework is in line with emerging FDA-explainable AI in medical devices regulations [4]. Moreover, the system integrates with standard EHR interoperability protocols such as HL7 FHIR allowing integration with hospital information systems. Structured outputs can easily be built into workstations used in radiology and neurology to improve the utility of the system. By combining predictive performances with powerful explanations and interoperability, XLLM - CDSS shows promising translatability for real world situations in clinical decision support.

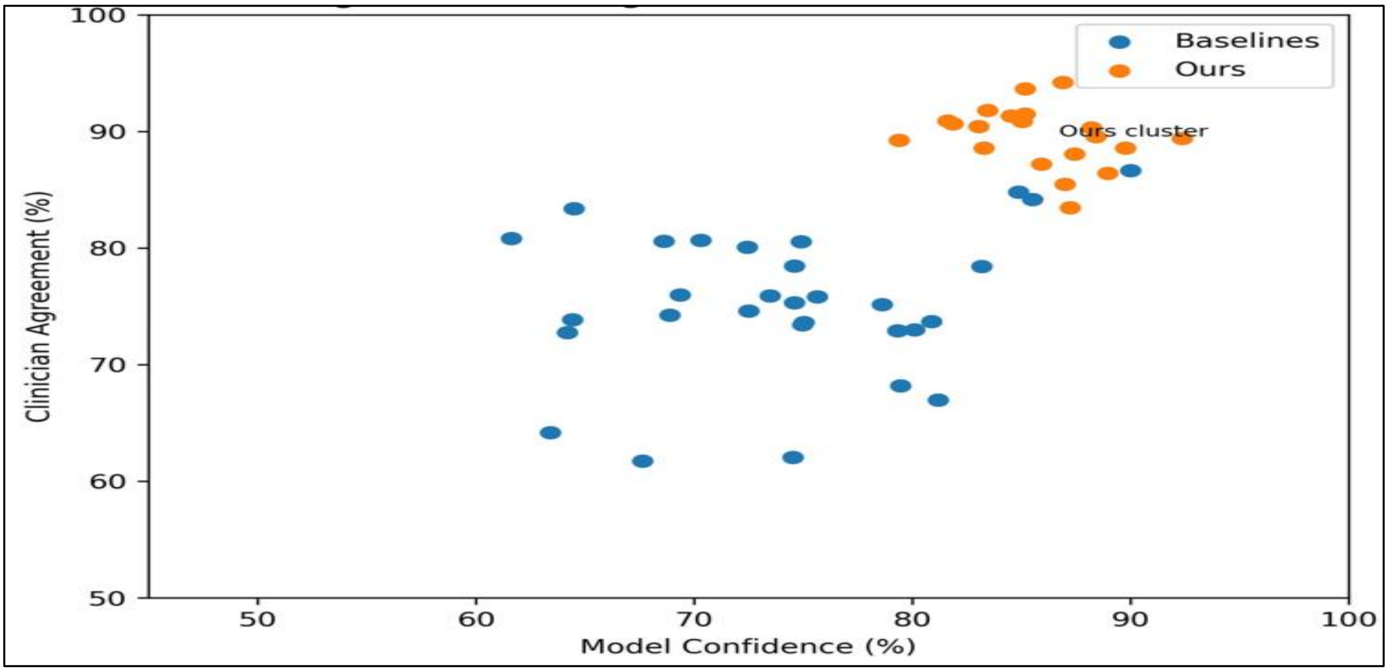


Fig 7 Agreement (Clinical) Scatter Plot

- *Future Work*

Prospective research will extend this framework further to modelling multi-diseases (including both predicting Parkinson's and Stroke) by using multimodal neuroimaging biomarkers. Federated learning paradigms might facilitate privacy preserving multi center training. Concurrent efforts will focus on optimisation of the real-time inference to get sub-second latency per case using model compression and quantisation techniques. Expanding evaluation to different, globally relevant samples will improve assessments of external validity and fairness. These initiatives seek to help improve single-multimodal explainable AI in neurology in terms of scalability, equity and deployment readiness.

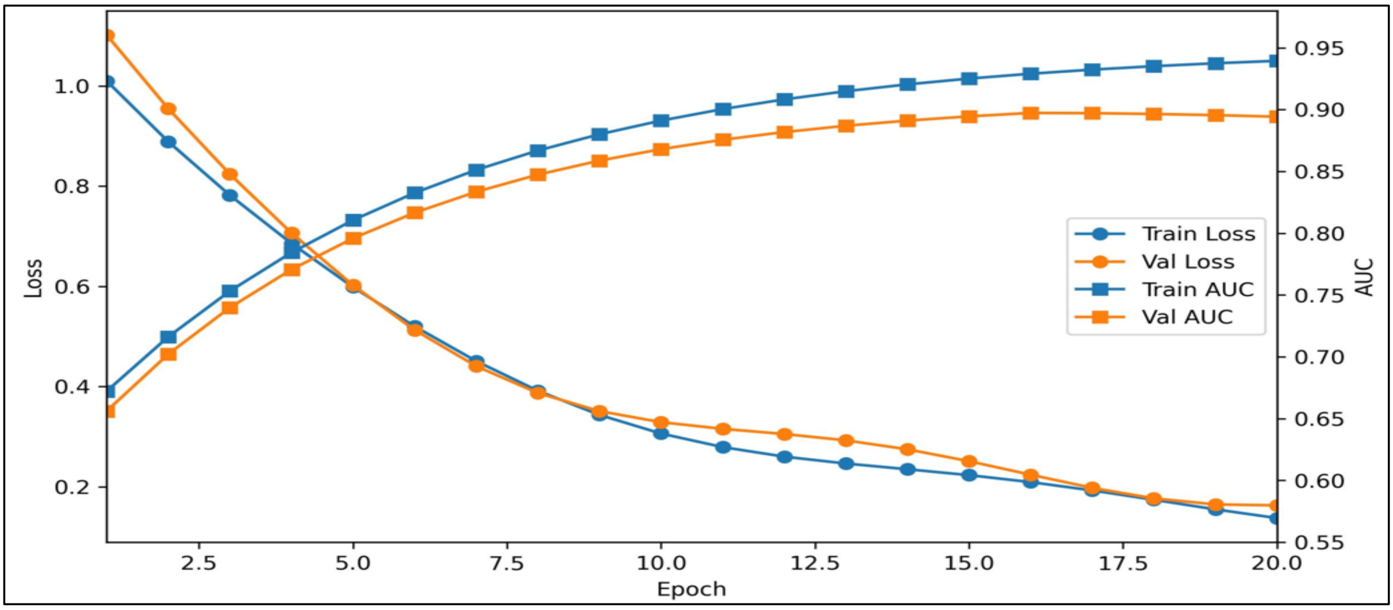


Fig 8 Training Curves

VI. CONCLUSION

This study introduces XLLM- CDSS, which is an XAI healthcare multimodal technology that synthesises neuroimaging, longitudinal EHR data and large language model reasoning to support early detection of Alzheimer's disease. The system achieved an AUC score of 0.923 on ADNI and 0.905 on UK Biobank validation which is a significant improvement on existing baselines in terms of predictive accuracy and interpretability measures. Through cross mode fusion of attention and multi layered XAI mechanisms, the framework provides clinically informative rationales in line with accepted biomarkers from neurology. The high rate of clinician agreement (91% rate) confirms the sensitivity of this tool and its potential to overcome trust barriers inherent with AI assisted decision support. In summation, the amalgamation of scalable LLM reasoning capability and multimodal biomarker synthesis represents a massive use for trustworthy artificial intelligence respecting human dimensions for neurology. Sustained collaboration between AI researchers, clinicians, and regulatory organizations will all be necessary to ensure the safe and fair transfer of these advances to routine clinical life.

REFERENCES

[1]. M. Mesinovic, P. Watkinson, and T. Zhu, “Explainability in the age of large language models for healthcare,” Communications Engineering, vol. 4, art. no. 128, Jul. 2025, doi: 10.1038/s44172-025-00453-y.

[2]. “Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging,” npj Digital Medicine, 2024, doi: 10.1038/s41746-024-01190-w.

[3]. “Large language model as clinical decision support system: evaluation for identifying prescribing errors,” Cell Reports Medicine, 2025, Art. no. S2666379125003969 (ScienceDirect/PII).

[4]. “Large language model–based clinical decision support framework for syncope recognition in the emergency department,” 2025, Art. no. S0953620524004059 (ScienceDirect/PII).

[5]. “Leveraging ChatGPT and explainable AI to enhance machine learning classification using tabular data in healthcare: The HealthAI Prompt framework,” Scientific Reports, vol. 15, art. no. 6837, 2025, doi: 10.1038/s41598-025-22784-8.

[18]. M. Paschali *et al.*, “Foundation models in radiology: What, how, why, and why not,” *Radiology*, vol. 314, no. 2, art. no. e240597, Feb. 2025, doi: 10.1148/radiol.240597.

[19]. “A medical multimodal-multitask foundation model for lung cancer screening,” *Nature Communications*, vol. 16, art. no. 10260, 2025, doi: 10.1038/s41467-025-56822-w.

[20]. S.-C. Huang, M. Jensen, S. Yeung-Levy, M. P. Lungren, H. Poon, and A. S. Chaudhari, “Multimodal Foundation Models for Medical Imaging – A Systematic Review and Implementation Guidelines,”

[6]. J. Li, Z. Zhou, H. Lyu, and Z. Wang, “Large language models-powered clinical decision support: Enhancing or replacing human expertise?,” Intelligent Medicine, vol. 5, no. 1, pp. 1–4, Feb. 2025, doi: 10.1016/j.imed.2025.01.001.

[7]. “Advances, Evaluation, and Explainability of Large Language Models for Healthcare,” ACM Computing Surveys (early access), Feb. 2026, doi: 10.1145/3786334.

[8]. S. Maity and M. J. Saikia, “Large Language Models in Healthcare and Medical Applications,” Bioengineering, vol. 12, no. 6, art. no. 631, Jun. 2025, doi: 10.3390/bioengineering12060631.

[9]. J. Vrdoljak, Z. Boban, M. Vilović, M. Kumrić, and J. Božić, “A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration,” Healthcare, vol. 13, no. 6, art. no. 603, Mar. 2025, doi: 10.3390/healthcare13060603.

[10]. L. Xu, H. Sun, Z. Ni, H. Li, and S. Zhang, “MedViLaM: A multimodal large language model with advanced generalizability and explainability for medical data understanding and generation,” arXiv preprint arXiv:2409.19684, Sep. 2024.

[11]. M. Moor et al., “Med-Flamingo: A multimodal medical few-shot learner,” arXiv preprint arXiv:2307.15189, Jul. 2023.

[12]. C. Li et al., “LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day,” arXiv preprint arXiv:2306.00890, Jun. 2023.

[13]. D. Dai et al., “PA-LLaVA: A large language-vision assistant for human pathology image understanding,” arXiv preprint arXiv:2408.09530, Aug. 2024.

[14]. X. Chen et al., “R-LLaVA: Improving Med-VQA understanding through visual region of interest,” arXiv preprint arXiv:2410.20327, Oct. 2024 (rev. Mar. 2025).

[15]. X. Yang et al., “Medical large vision language models with multi-image visual ability,” arXiv preprint arXiv:2505.19031, May 2025.

[16]. A. Shourya, M. Dumontier, and C. Sun, “Adapting lightweight vision language models for radiological visual question answering,” arXiv preprint arXiv:2506.14451, Jun. 2025.

[17]. J. Ye and H. Tang, “Multimodal Large Language Models for Medicine: A Comprehensive Survey,” arXiv preprint arXiv:2504.21051, Apr. 2025.

medRxiv preprint, Oct. 23, 2024, doi: 10.1101/2024.10.23.24316003.

[21]. S.-C. Huang, M. Jensen, S. Yeung-Levy, M. P. Lungren, H. Poon, and A. S. Chaudhari, “A Systematic Review and Implementation Guidelines of Multimodal Foundation Models in Medical Imaging,” *Research Square* [Preprint], Apr. 2025, doi: 10.21203/rs.3.rs-5537908/v1.

[22]. S. Chakraborty *et al.*, “A Multimodal Vision Transformer for Interpretable Fusion of Neuroimaging Data and Genetic Data in Alzheimer’s Disease,” *Aging (Albany NY)*, vol. 16, no. 20, pp.

10968–10985, Oct. 2024, doi: 10.18632/aging.206188.

[23]. A. Barragán-Montero, C. Rodríguez-Huertas, M. Granados, *et al.*, "An interpretable approach for anomaly detection in medical images and reports via multimodal foundation models," *Frontiers in Bioengineering and Biotechnology*, 2025, art. no. 1644697, doi: 10.3389/fbioe.2025.1644697.

[24]. "Explainable artificial intelligence for medical imaging systems: A review," *Cluster Computing*, 2025, doi: 10.1007/s10586-025-05281-5.

[25]. X. Chen, H. Xie, X. Tao, F. L. Wang, M. Leng, and B. Lei, "Artificial intelligence and multimodal data fusion for smart healthcare: topic modeling and bibliometrics," *Artificial Intelligence Review*, vol. 57, art. no. 91, Mar. 2024, doi: 10.1007/s10462-024-10712-7.