Cloud Computing and Data Science
https://ojs.wiserpub.com/index.php/CCDS/

UNIVERSAL WISER PUBLISHER

Research Article

Advancing Explainable AI for Clinical Decision Support: A Multimodal Evaluation Framework

Hashim Ali

Department of Computer Science, Nazarbayev University, Astana, 010000, Kazakhstan
E-mail: hashim.ali@nu.edu.kz

Received: 28 February 2026; **Revised:** 14 April 2026; **Accepted:** 14 April 2026

Abstract: The increasing integration of Artificial Intelligence (AI) into Clinical Decision Support Systems (CDSS) is constrained by the limited transparency of model predictions, which undermines clinician trust and slows adoption in safety-critical settings. To address this barrier, we propose a unified multimodal evaluation framework for eXplainable AI (XAI) and empirically assess the behavior of the explanations in four clinically relevant modalities: chest radiography for pathology detection, Electroencephalography (EEG)-based epilepsy decision support using High-Frequency Oscillation (HFO) evidence, multimodal emotion recognition for psychological decision support and prediction of Alzheimer's disease based on Electronic Health Records (EHR). The framework evaluates widely used explanation mechanisms, including Gradient-weighted Class Activation Mapping (Grad-CAM), Integrated Gradients (IG), SHapley Additive exPlanations (SHAP)-style feature attribution, and attention-based interpretation, using modality-appropriate criteria that emphasize reliability, robustness, and clinical plausibility rather than visualization quality alone. The results show that Grad-CAM provides stable region-level localization in chest X-ray prediction. In contrast, EEG-based epilepsy decision support requires interpretability grounded in domain-specific biomarkers and time-frequency structure rather than generic saliency. In multimodal emotion recognition, fusion improves performance, but the contribution of each modality varies by emotional state, highlighting the need for interpretable fusion analyses. For Alzheimer's prediction, a tuned CatBoost model achieves strong discrimination, and feature-level analyses identify clinically significant drivers, including Mini-Mental State Examination (MMSE)-related measures. Cross-modal synthesis demonstrates that explanation effectiveness is inherently task- and modality-dependent and that explanation instability and susceptibility to spurious cues remain recurring risks across settings, consistent with concerns about the faithfulness of saliency explanations. In general, the proposed framework supports the practical implementation of XAI in healthcare by providing modality-aligned guidance for selecting and validating explanations within clinical workflows.

Keywords: explainable artificial intelligence, clinical decision support system, multimodal learning, medical imaging, Electroencephalography (EEG), electronic health records, robustness

1. Introduction

Artificial Intelligence (AI)-enabled Clinical Decision Support Systems (CDSS) have demonstrated potential to improve diagnostic precision, risk stratification, and personalized treatment planning in multiple specialties [1,

DOI: https://doi.org/10.37256/ccds.7220269831

2]. However, many state-of-the-art models remain opaque, creating a barrier to adoption in clinical practice where accountability, transparency, and patient safety are central requirements [3]. This challenge is especially acute in high-stakes settings: clinicians can reject a correct prediction without a credible rationale. In contrast, an incorrect prediction accompanied by a persuasive but misleading explanation can introduce new risks to patient care [3, 4]. Consequently, interpretability is increasingly viewed not as an optional feature but as an enabling condition for trustworthy AI deployment in healthcare.

EXplainable AI (XAI) has emerged as a family of approaches aimed at increasing transparency by providing human-interpretable descriptions of model behavior, such as saliency maps, feature attributions, counterfactuals, or attention-based relevance scores [5]. However, the reliability and clinical usefulness of XAI remain inconsistent and modality dependent. Techniques that appear effective in one domain can degrade when applied to different types of data or clinical tasks [2, 5]. For example, gradient-based localization methods (e.g., Gradient-weighted Class Activation Mapping (Grad-CAM)) can produce intuitive region-level explanations for imaging models [6]. However, analogous saliency-based explanations may be insufficient for Electroencephalography (EEG) signals, which are temporally and spectrally structured, often linked to High-Frequency Oscillation (HFO)-related biomarkers and timing effects [7]. Similarly, for the tabular prediction-style Electronic Health Record (EHR), the explanations must align with interpretable clinical variables and remain robust to imbalances and correlated predictors, which can distort apparent importance [5, 8].

Despite rapid methodological progress, much of the evidence base for XAI in healthcare is fragmented. Existing studies often evaluate explanations within a single modality or clinical domain, limiting their generalizability to real-world CDSS ecosystems that increasingly integrate imaging, biosignals, EHRs, and multimodal behavioral data [2, 9, 10]. Furthermore, evaluation protocols commonly emphasize technical fidelity while underweighting clinical plausibility and workflow relevance, even though clinician trust depends heavily on whether explanations align with domain reasoning and remain stable under clinically irrelevant perturbations [2, 5, 11].

This study introduces a unified multimodal evaluation framework. It empirically assesses XAI across four representative CDSS settings: (i) chest X-ray pathology detection (spatial evidence), (ii) EEG/HFO-guided epilepsy decision support (temporal and time-frequency evidence) [7], (iii) multimodal emotion recognition for psychological support (fusion and contribution stability) [2, 12], and (iv) EHR-based Alzheimer's disease prediction (variable-level evidence). The framework evaluates explanations along the modality-aligned axes, faithfulness/reliability, robustness/stability, and clinical plausibility/workflow fit, consistent with responsible CDSS guidance [2] and medically oriented XAI design principles [11].

The main contributions are threefold: (i) a consolidated cross-modal evaluation of explanation mechanisms spanning spatial imaging, temporal biosignals, multimodal behavioral inference, and structured EHR prediction [2, 5]; (ii) identification of recurring failure modes, including instability, spurious cue sensitivity, and modality mismatch, that emerge across domains despite differing data types and model families [4]; and (iii) actionable, modality-specific guidance for selecting and validating explanations in clinical workflows, positioning explainability as a validated system property rather than a visualization artifact [3].

The remainder of this paper is structured as follows. Section 2 reviews related work on XAI in healthcare and highlights limitations of current evaluation practices. Section 3 describes the proposed framework, datasets, models, and explanation methods. Section 4 reports results and cross-modal trends. Section 5 discusses implications, limitations, and deployment guidance. Section 6 concludes with recommendations for future research and practical implementation.

2. Review of the literature

XAI has become a central research topic in AI in healthcare, reflecting the growing recognition that interpretability is critical to accountability, safety, and clinical adoption [3, 5]. Contemporary surveys organize XAI methods into intrinsic vs. post-hoc explanations, local vs. global explanations, and modality alignment (images, signals, text, tabular data) [5]. Across clinical domains, two patterns recur: explanation mechanisms are often selected based on model class rather than clinical need, and explanation evaluation is less standardized than performance evaluation, limiting clinical conclusions [3, 5].

In radiology and related imaging applications, gradient-based explanation methods, such as Grad-CAM and Integrated Gradients (IG), are widely used to interpret deep convolutional models [6, 13]. Heatmaps are intuitive and align with clinicians' expectations for spatial localization. However, saliency maps can highlight regions unrelated to pathology, react to spurious artifacts, and exhibit limited faithfulness under perturbations, issues that become especially problematic in multi-label and complex cases [4, 14]. These concerns motivate evaluating beyond qualitative plausibility, including robustness and faithfulness tests [4].

The interpretability of EEG differs fundamentally from that of imaging, because clinically meaningful evidence is often temporal and spectral rather than spatial. In epilepsy decision support, HFOs and timing effects are increasingly recognized as relevant for identifying epileptogenic tissue, and explanations must align with these domain markers to be clinically meaningful [7]. The broader neuropsychiatric EEG literature similarly emphasizes biomarker grounding and interpretability-aware methodology, rather than relying exclusively on generic attribution [15]. Therefore, clinically oriented XAI design has advocated the integration of domain constraints and medically meaningful representations to enhance interpretive validity [11].

Multimodal AI introduces additional complexity because explanations must address not only why a prediction was made, but also how evidence from heterogeneous modalities was fused. Reviews of multimodal medical diagnosis highlight both the promise and complexity of combining modalities, including the need for consistent interpretability across heterogeneous data streams [2]. In mental health applications that integrate EEG and other modalities, systematic evidence indicates the importance of modality-aware interpretation and the limitations of purely technical attention scores when they do not align with clinical constructs [11, 12]. Multimodal CDSS efforts that integrate predictive modeling, imaging biomarkers, and EHR-linked decision support further underscore the need for transparent explanations that align with clinical workflows and data integration realities [10].

EHR-based prediction has historically benefited from interpretable baselines, but contemporary CDSS increasingly rely on ensembles and other high-capacity learners that require post hoc explanations [1, 3, 5]. SHAP-style feature attribution is widely adopted for tabular clinical data because it supports both local and global interpretation and works well with tree-based models; surveys on clinical risk prediction discuss its adoption and limitations [5]. However, neuro-symbolic approaches and other paradigms emphasize that interpretability must align with clinical reasoning and avoid confounding proxies, sampling artifacts, and handling of imbalances [8]. As a result, stability checks across folds and sampling strategies are often considered essential to avoid overconfident interpretations [3, 5].

Cloud computing and data science platforms increasingly underpin clinical AI pipelines by enabling scalable storage, high-throughput preprocessing, and distributed model training for large, heterogeneous datasets. From an XAI perspective, scalability becomes a first-order constraint: explanation methods that are feasible on small benchmarks may be impractical for large-scale clinical data or real-time CDSS deployments, particularly when explanations are needed for many patients, across multiple modalities, or with frequent model updates. Recent work on scalable, distributed AI frameworks describes cloud-native orchestration patterns that reduce end-to-end latency and cost for deep learning workloads, which becomes relevant when explanation computation is integrated into CDSS inference pipelines [16].

Parallel and distributed explainability has therefore emerged as an active area of research. For example, Shapley-style attribution has been implemented on distributed data science systems to increase throughput for large models and feature spaces. In privacy-preserving and multi-party settings, explanation consistency is itself a challenge; distributed KernelSHAP-style approaches have been proposed to produce explanations that remain comparable across collaborators. In healthcare, where multi-center data sharing is constrained, federated learning is increasingly used, and recent studies and surveys highlight the need for explainability mechanisms compatible with federated workflows (e.g., explanation aggregation, stability monitoring, and privacy-aware interpretation) [17-20].

Three gaps recur in the literature. First, the evidence is predominantly single-modal, providing limited support for method selection in multi-source clinical environments where imaging, biosignals, and EHR routinely coexist [2, 9, 10]. Second, evaluation protocols often emphasize technical fidelity while giving insufficient weight to clinical plausibility and workflow constraints, even though these are central to clinician trust [3, 11]. Third, recurring failure modes; instability, spurious cues, and limited faithfulness, are widely discussed, but are less often documented systematically in diverse tasks within a single framework [4].

Existing literature surveys and reviews organize methods and report modality-specific findings [3, 9], while clinically oriented frameworks emphasize medically grounded design [11]. However, few studies consolidate evidence

side by side across spatial (imaging), temporal (EEG), fusion (multimodal behavior), and variable-based (EHR) settings, using a consistent set of explanation-quality axes directly relevant to the adoption of CDSS [3, 10]. This paper addresses that gap by evaluating explanation behavior across modalities under a unified rubric (faithfulness, robustness, clinical plausibility) [3, 4], identifying recurring cross-domain failure modes, and deriving actionable deployment guidance.

3. Materials and methods

The presented study consolidates four independent healthcare-oriented intelligent system projects into a unified explainable clinical decision support framework, as shown in Figure 1: (i) acquisition and preprocessing of data per modality, (ii) development and training of models, (iii) generation of explainability (attribution/attention/importance of features/clinical rule outputs), and (iv) evaluation of modality-specific and cross-modal explanations. Rather than treating each system as an isolated technical contribution, this study adopts a modality-spanning perspective in which predictive modeling and explainability are evaluated under a common methodological lens. Four clinical modalities are linked to such clinical systems, which include: (i) classification of multi-label chest radiology (Chest X-ray), (ii) epilepsy localization support using High-Frequency Oscillation (HFO) analysis from EEG signals, (iii) multimodal emotion recognition to support mental-health assessment, and (iv) prediction of Alzheimer's disease from structured EHRs. The central objective is not merely to demonstrate high predictive performance, but to evaluate how explanation techniques behave across fundamentally different clinical data modalities and whether those explanations are sufficiently robust, faithful, and clinically plausible to support real-world deployment.

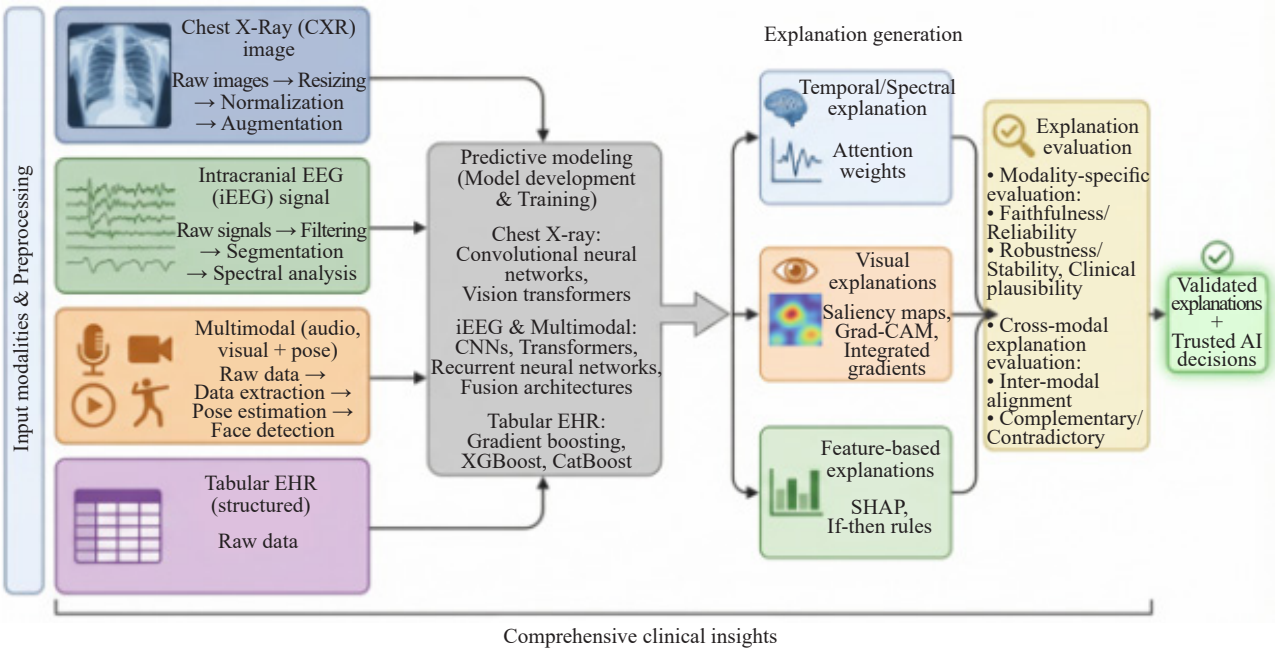


Figure 1. Unified multimodal XAI evaluation workflow framework

3.1 *Classification and explanation of chest radiograph pathology*

The CheXpert chest radiograph benchmark [21], released by Stanford University, comprises 224,316 frontal-view chest X-ray images from 65,240 patients, annotated for 14 common radiologic findings, including uncertainty labels. The Chest X-ray pathology pipeline is shown in Figure 2. Uncertainty labels were handled using the U-ones strategy (treating uncertainty as positive), consistent with the original dataset's recommendations. The study followed the

original benchmark split, using approximately 223 k images for training, 200 images for validation, and 500 images for testing. All images were resized to 320 × 320 pixels, normalized, and augmented during training with random flips and rotations to improve generalization.

For pathology classification, the DenseNet-121 architecture was utilized, a 121-layer convolutional architecture widely used for medical image recognition due to dense connectivity and strong gradient propagation. The output layer was configured for 14 independent pathology scores with sigmoid activations, reflecting the multi-label nature of CheXpert. The model was trained using binary cross-entropy across the 14 labels (summed over outputs), optimized with Adam (initial learning rate 1×10^{-4}; $\beta_1 = 0.9$, $\beta_2 = 0.999$), for 5 epochs with batch size 64, saving checkpoints periodically (every 4,800 iterations).

To generate post-hoc visual explanations for DenseNet-121 predictions, four gradient-based attribution techniques were applied: (i) Vanilla Gradients (saliency maps), (ii) Grad-CAM [6], (iii) Integrated Gradients (IG) [13], and (iv) region-based (XRAI-style) explanations. Vanilla Gradients were computed as input-output derivatives, offering pixel-level sensitivity but often producing noisy, unstable maps in deep Convolutional Neural Networks (CNNs). Grad-CAM was used to localize influential regions via gradients with respect to the last convolutional feature maps, producing coarser but more clinically interpretable heatmaps (for simplicity, an enhanced localization variant is described as Grad-CAM). IG was computed by integrating gradients along a baseline-to-input path (with the baseline set to all zeros), using 50 interpolation points to reduce noise and saturation, yielding smoother attributions. XRAI was used to produce region-based attributions on segmented superpixel regions, with the aim of improving clinical coherence by emphasizing anatomically meaningful regions. Figure 3 presents a qualitative comparison of the explanation heatmaps for a CheXpert test image that was predicted positive for cardiomegaly (and pleural effusion).

The quality of the explanation was assessed using a structured qualitative protocol applied to representative test cases, comparing how each method highlighted the clinically significant anatomy of the target findings. For example, in cases predicted to have cardiomegaly and pleural effusion, saliency maps were generated for each method and visually analyzed for coherence, plausibility of localization, and noise/fragmentation characteristics. The evaluation explicitly contrasted pixel-noisy behaviors (often observed in Vanilla Gradients) with region-level interpretability and anatomical plausibility (often clearer in Grad-CAM/IG/XRAI), and documented failure modes such as clinically unrelated anatomical structures being merged into a single region under XRAI segmentation.

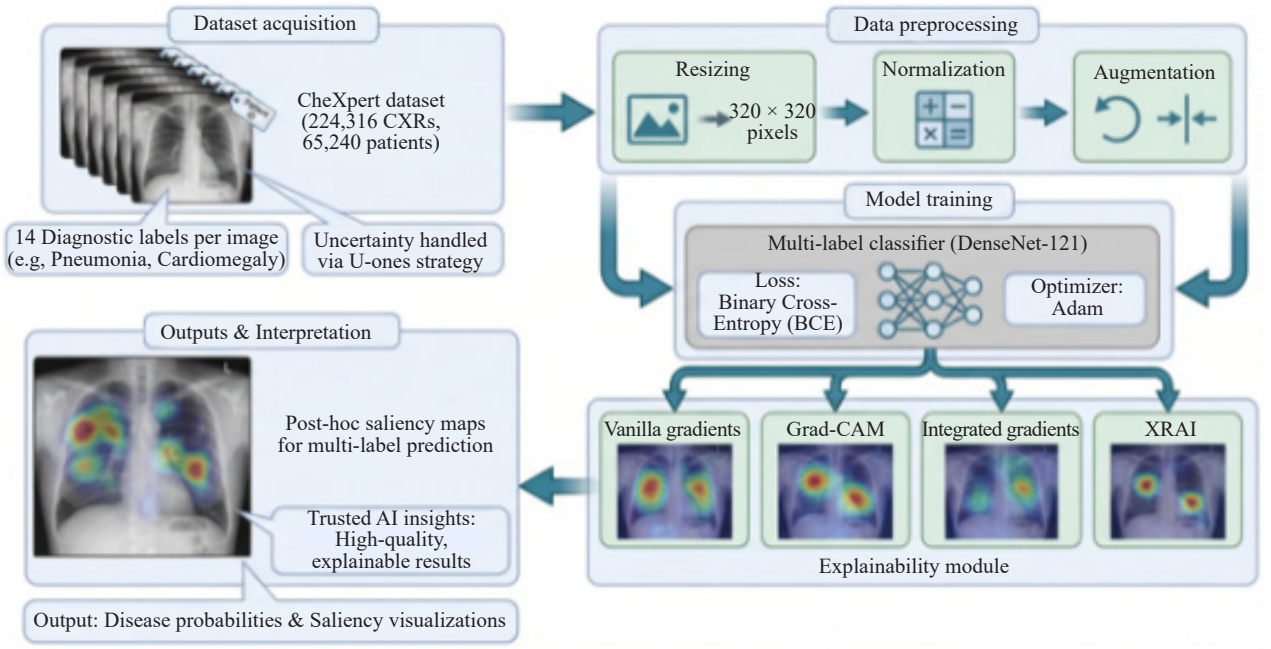


Figure 2. Chest X-ray pipeline for pathology detection and explainability

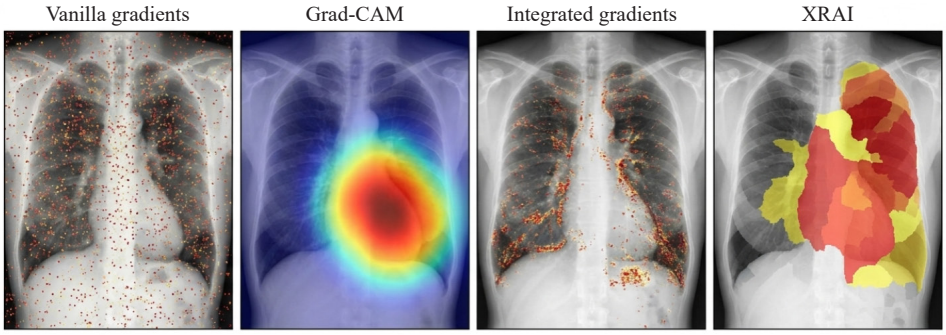


Figure 3. Chest X-ray pipeline for pathology detection and explainability, with a CheXpert test image predicted as positive for cardiomegaly (and pleural effusion). Saliency maps for cardiomegaly output are generated using Vanilla gradients, Grad-CAM, Integrated gradients, and XRAI and overlaid on the original radiograph. The panel illustrates method-specific differences in noise, spatial coherence, and anatomical plausibility

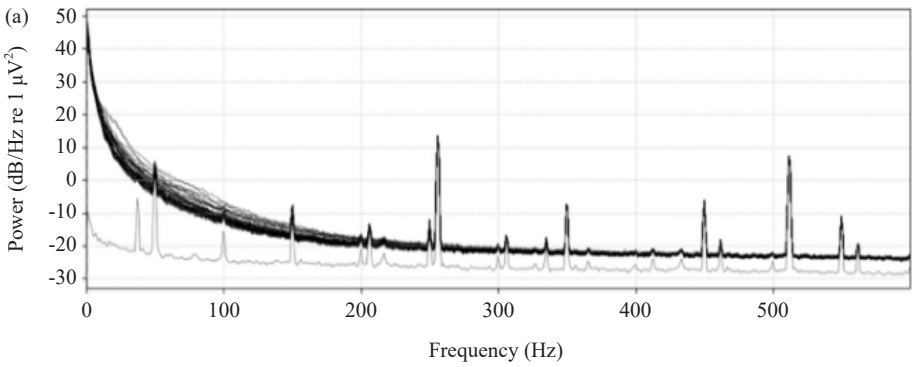
3.2 *EEG-based epilepsy decision support*

Intracranial EEG (iEEG) recordings [22] were preprocessed to isolate HFOs, with a focus on clinically relevant ripple-band activity. Line noise was attenuated using notch filtering, preserving the spectral profile of broadband $1/f$ while removing mains-frequency peaks, followed by band-pass filtering to retain 80-500 Hz activity (conventional HFO range). Signals were then re-referenced using a bipolar montage by subtracting adjacent contacts along each electrode, reducing common-mode artifacts and improving the spatial focus of detected HFO activity.

To construct supervised learning examples, 0.2 s windows centered on expert-marked events were extracted as positive HFO segments. In contrast, matched-duration background windows were sampled from non-annotated periods on the same channels/runs as non-HFO examples. Time-frequency representations were inspected using Morlet wavelet transforms to verify that positive windows exhibited concentrated power increases around 100-300 Hz at event onset, whereas non-HFO windows lacked comparable bursts; averaged maps further highlighted consistent ripple-band differences.

Two deep learning models were used for window-wise detection: a 1D CNN (HFONet1D) and a 1D Transformer (HFOTransformer1D). Training, validation, and test splits were constructed at the window level with large-scale sampling; the reported split sizes were N_{train} = 932,165, N_{val} = 485,744, and N_{test} = 446,940 windows, with approximately balanced class proportions (~ 50% each), enabled through the sampling strategy.

Interpretability was studied using gradient-based saliency profiles (for HFONet1D) and self-attention visualization (for HFOTransformer1D). For CNN, mean saliency was computed separately for correctly classified HFO and non-HFO windows, revealing a dominant peak near the center of the 0.2 s window (time 0) with a falloff toward the edges, consistent with decisions driven by local oscillatory morphology around the event. For the Transformer, attention maps from an early encoder layer exhibited a strong diagonal pattern (emphasizing local temporal neighborhoods) and concentrated weights near the middle of the window, aligning with short HFO durations. Figures 4-6 illustrate spectral preprocessing and representative time-frequency signatures that support biomarker-based interpretation.



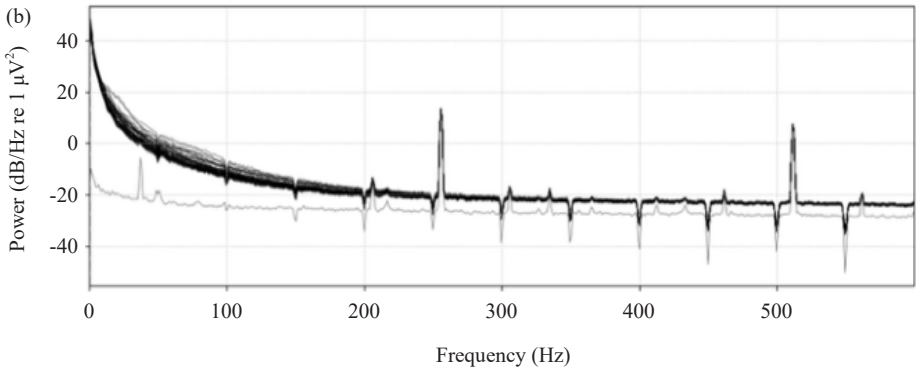


Figure 4. Power Spectral Density (PSD) of raw intracranial EEG before and after notch filtering. Notch filtering attenuates mains-frequency peaks and harmonics while preserving the broadband 1/*f* structure before band-pass filtering for HFO analysis

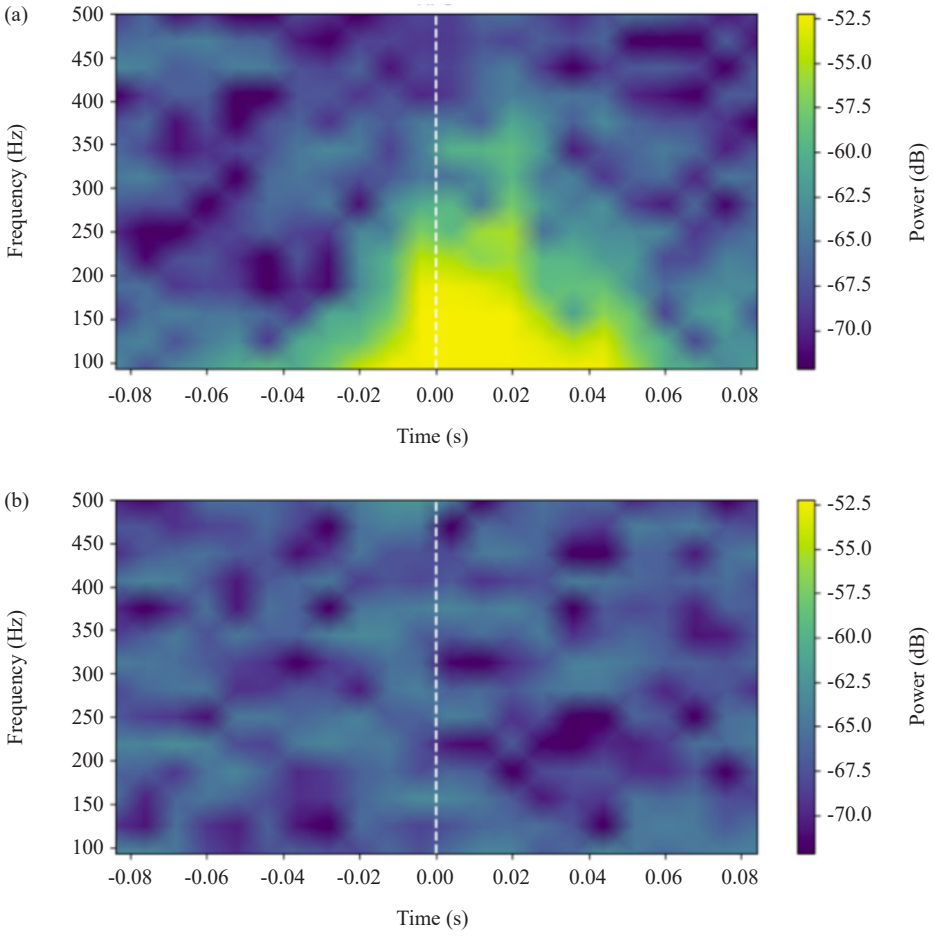


Figure 5. Morlet wavelet time-frequency maps for (a) an HFO-positive window and (b) a non-HFO window. HFO windows show a time-locked burst of power typically within ~ 100-300 Hz around the onset of the event, whereas non-HFO windows lack a comparable high-frequency burst

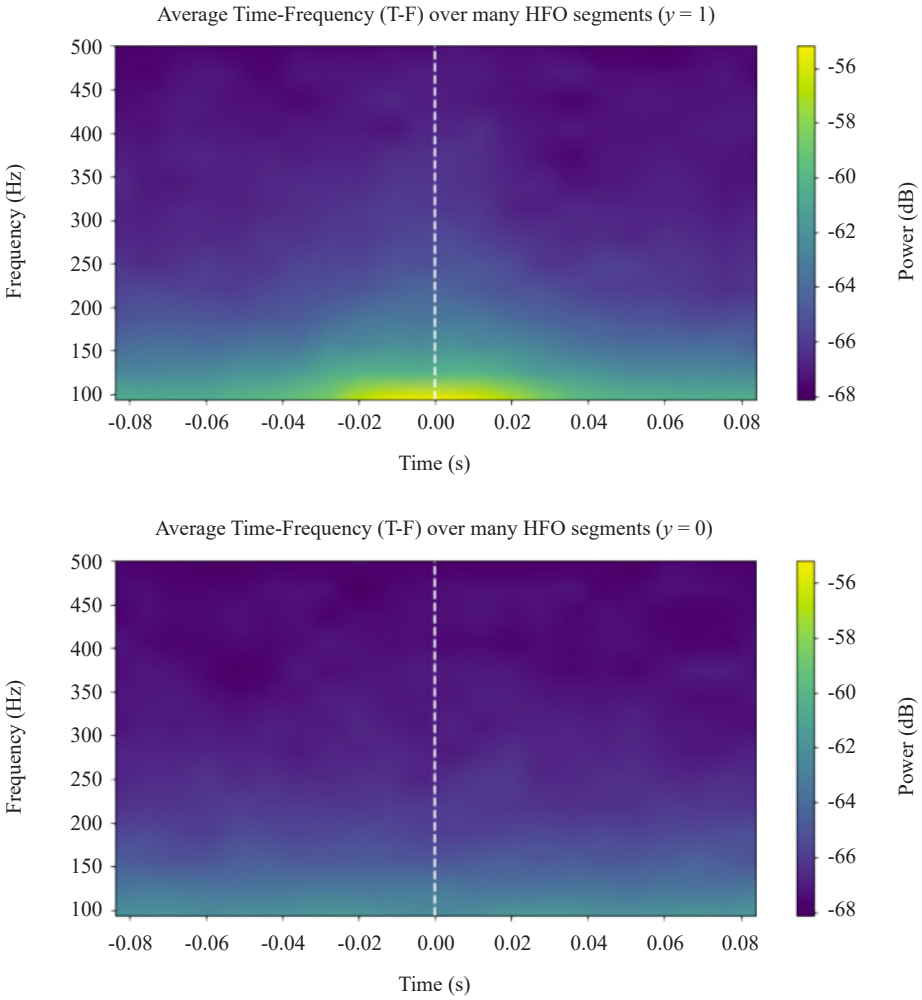


Figure 6. Mean Time-Frequency (T-F) maps averaged across many segments highlight a consistent increase in ripple-band power that is time-locked to HFO windows, while non-HFO windows remain comparatively flat

3.3 *Multimodal emotion recognition for psychological decision support*

A trimodal setup was constructed using the CREMA-D dataset [23], with synchronized video, audio, and pose representations stored per sample as NumPy tensors, as shown in Figure 7. Each sample consisted of (i) video features of shape 90 × 1,408, (ii) audio features of shape 50 × 1,024, (iii) pose features of shape 90 × 156, plus a label file. The pose sequences were repaired for missing frames using a deterministic imputation strategy: missing initial frames were filled forward, trailing missing frames were filled back, short gaps (≤ 8 frames) were linearly interpolated, and longer gaps were filled by carrying the most recent valid pose, yielding stable sequences of fixed length 90 × 156.

For normalization, the mean and standard deviation for each feature were estimated from the training set only (up to 1,000 samples, per modality), and the same statistics were applied to the validation/test sets to prevent data leakage. Each modality sequence was constrained to a fixed target length via center cropping (if longer) or zero-padding (if shorter). Robust evaluation used five independent random splits, with 70% training, 15% validation, and 15% test per split; models were trained independently across splits, with each split serving as a cross-validation fold.

The core model was a trimodal neural network built around E-Branchformer encoders. Modality-specific encoders projected each input to a shared hidden size of 2,048, processed it through E-Branchformer layers (8 attention heads), and mapped the output to a latent dimension of 50 per timestep, preserving the original sequence lengths (90 for video and pose; 150 for audio). The temporal mean pooling converted each modality sequence into a 50-dimensional vector, and the concatenation produced a fused vector of 150 dimensions.

A shared fusion encoder then modeled cross-modal interactions by reshaping the fused vector into a sequence of length-1 (batch × 1 × 150), projected to 2,048, applying two layers of the E-Branchformer, and projecting to a shared representation of 512 dimensions for downstream heads. The classification head mapped 512 → 1,536 → 512 → 6 logits, using Sigmoid Linear Unit (SiLU) activations and dropout ($p = 0.25$).

Two complementary interpretability mechanisms were integrated into the model design. First, attention-based interpretability is naturally supported by the E-Branchformer's multi-head self-attention blocks in each modality encoder and in the shared fusion encoder, enabling analysis of temporal salience and cross-modal interactions via attention patterns (e.g., which frames/segments dominate the fused representation). Second, an auxiliary reconstruction head encouraged the shared representation to preserve information from all modalities: the shared 512-dimensional vector was projected to a normalized 50-dimensional reconstruction vector, trained against a target built by mean-pooling the raw modality inputs and projecting them to a normalized 50-dimensional target. This design supports explanation by probing which components of each modality are retained in the shared representation and by analyzing reconstruction alignment across modalities.

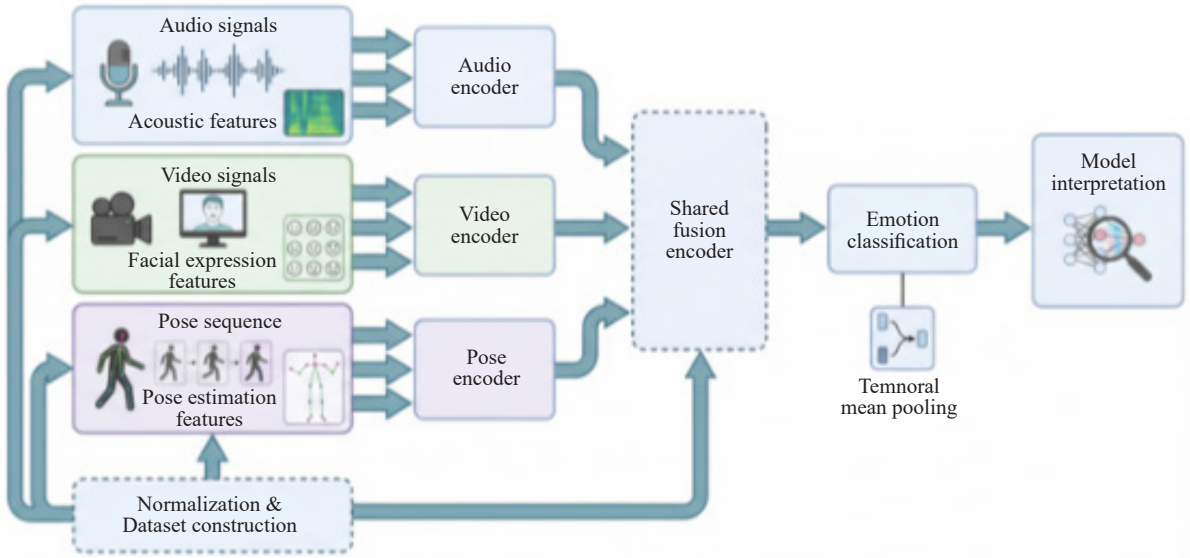


Figure 7. Systeme of fusion architecture for multimodal emotion recognition and contribution analysis

3.4 *EHR-based Alzheimer's disease prediction*

A publicly available Kaggle Alzheimer's Disease Dataset [24] was used, which contains 2,149 patient records with a binary diagnosis label (Class 0: not diagnosed; Class 1: diagnosed). Exploratory analysis indicated moderate class imbalance (~ 65% Class 0 vs. ~ 35% Class 1), motivating imbalance-aware resampling and evaluation.

Data cleaning and feature preparation were performed using pandas, while categorical attributes (e.g., Smoking, Gender) were encoded as numeric features using scikit-learn utilities to enable downstream machine learning. Outliers were removed using an Interquartile Range (IQR)-based procedure applied to 11 continuous numerical variables to reduce the influence of anomalous values on training. To address class imbalance without leakage, Synthetic Minority Oversampling Technique (SMOTE) oversampling was incorporated into the modeling pipeline, so that oversampling occurred only in the training portion of each cross-validation fold.

Model selection compared eight algorithms spanning classical and modern ensemble learners, including Gradient Boosting, eXtreme Gradient Boosting (XGBoost), and CatBoost, and evaluated them using 10-fold stratified cross-validation to assess generalization in the presence of class imbalance. The macro F1-score was used as the primary selection metric because it is robust to skewed class distributions. CatBoost was selected as the best-performing model, then tuned using GridSearchCV and evaluated on a holdout test set.

Interpretability was operationalized by analyzing the importance of features of the final CatBoost model to support clinical plausibility checks. The most influential predictors were clinically significant cognitive/functional measures: Functional Assessment, Activity of Daily Living (ADL), Memory Complaints, Mini-Mental State Examination (MMSE), and Behavioral Problems, supporting the hypothesis that predictions were driven by substantive clinical factors rather than spurious correlations. Correlation analysis (Pearson) complemented this by summarizing linear relationships between features and diagnosis, while acknowledging that many predictors may contribute via non-linear interactions better captured by ensemble models. Figure 8 outlines the pipeline.

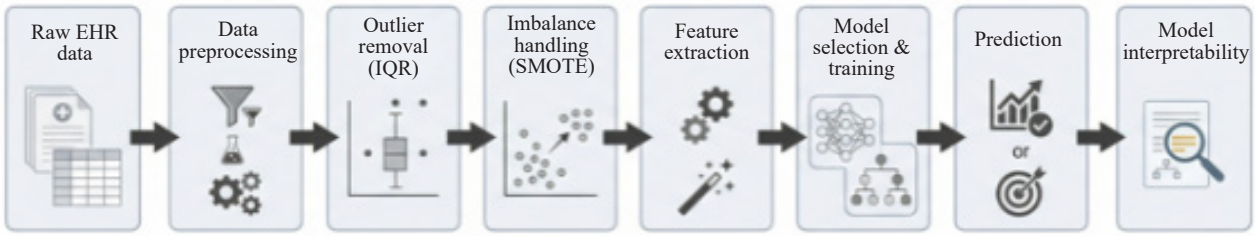


Figure 8. EHR-based Alzheimer prediction workflow: removal of outliers using IQR on continuous variables, handling of imbalances using SMOTE within a leakage-safe Machine Learning (ML) pipeline, and selection of models through 10-fold stratified cross-validation using Macro-F1; the final interpretability of the model is summarized using CatBoost's importance of features

3.5 *Cross-modal explainability evaluation strategy*

To make the four decision-support tasks comparable under a single evaluation framework, a cross-modal strategy is proposed that separates (i) predictive validity from (ii) explanation validity and then harmonizes explanation scoring through modality-appropriate proxies. Table 1 summarizes tasks, model families, and explanation mechanisms.

Across all modalities, model performance should be reported using task-appropriate discrimination metrics (e.g., Receiver Operating Characteristic-Area Under Curve (ROC-AUC) and F1 for binary tasks; Macro-F1 for imbalanced settings; subject-wise reporting for EEG to capture inter-patient variability). The EEG pipeline already supports this explicitly through window-wise and per-subject metrics, as well as ROC-AUC reporting. For Alzheimer prediction, Macro-F1 under stratified Cross-Validation (CV) and holdout tests is central due to class imbalance and asymmetric clinical costs.

Table 1. Unified overview of tasks, model families, and explanation mechanisms

Modality	Clinical task	Model family	Explanation mechanisms
Chest X-ray	Pathology detection	DenseNet-121	Grad-CAM; IG; region-based; gradients
EEG	Epilepsy decision support (HFO evidence)	Biomarker-Based Modeling	Time-frequency evidence + temporal attribution
Multimodal	Emotion recognition for psychological support	Fusion network	Attention + modality contribution + ablation validation
EHR	Alzheimer prediction	CatBoost	SHAP-style attribution + stability considerations

For explainability, the framework should align each modality with a pair of faithfulness and coherence criteria.

• Faithfulness (model-based): whether the highlighted evidence is causally/operationally important to the model's output. For X-ray and EEG, gradient-based sensitivity/attribution methods naturally support such assessments (e.g., saliency profiles, attention maps), and the methodology already includes a structured qualitative comparison and

documentation of known instability modes (e.g., noisy gradients; region-merging artifacts).

• Clinical coherence (human-based): whether explanations align with domain expectations (anatomically plausible localization in X-ray; temporally localized HFO morphology in EEG; clinically meaningful cognitive/functional drivers in EHR; and cross-modal consistency signals in emotion recognition through attention patterns and reconstruction alignment).

The implementation overview of the explainability methods used in this study is presented below; detailed information is provided in the Appendix.

Grad-CAM for Chest X-Ray (CXR): Grad-CAM explanations were generated with respect to the target pathology logit (pre-sigmoid) for each label. The target layer was selected as the last convolutional block of DenseNet-121 (i.e., the final feature map before global pooling), consistent with the Grad-CAM requirement for spatially meaningful convolutional activations. The heat maps were upsampled to the input resolution using bilinear interpolation and normalized to [0, 1] per image for visualization and quantitative scoring.

Integrated Gradients (CXR): IG used an all-zero baseline image and 50 interpolation steps between baseline and input, which matches the configuration already described in the manuscript. Attributions were aggregated into absolute per-pixel values and normalized to a unit-sum map to ensure comparability across samples.

Region-based explanations (XRAI-style, CXR): The images were partitioned into superpixel regions using Simple Linear Iterative Clustering (SLIC) (~ 300 regions per image). The region salience was calculated by aggregating the IG attribution values within each region. The regions were ranked by aggregate attribution and progressively added until a target coverage threshold, producing a region-level mask intended to improve anatomical coherence.

Attention-based interpretation (EEG and multimodal fusion): Attention weights were taken after softmax normalization within each attention head. For reporting, attention maps were averaged across heads and/or layers, as specified in the Appendix. For comparability across samples, attention matrices were normalized to unit row sums and summarized using temporal concentration metrics.

SHAP-style explanations (EHR/CatBoost): For tree-based CatBoost, SHAP values were computed using TreeSHAP (TreeExplainer) rather than KernelSHAP, because it is exact/efficient for trees and avoids kernel configuration. Background samples were drawn from the training set (see Appendix for details). Practical guidance on SHAP reporting and background selection follows established recommendations.

3.6 *Evaluation of the quantitative cross-modal explainability*

To enable systematic comparison across modalities, the quality of explanations was quantitatively assessed along three dimensions: faithfulness, robustness, and clinical plausibility. Faithfulness and robustness were evaluated employing modality-appropriate quantitative proxies; plausibility was assessed using a structured approach.

3.6.1 *Faithfulness (based on the model)*

Imaging (CXR): insertion/deletion AUC were computed by progressively inserting (or deleting) pixels in descending order of attribution magnitude and tracking the target logit/probability response; larger insertion AUC and smaller deletion AUC indicate more faithful explanations.

EEG: time-mask deletion was calculated by masking the top-attributed time indices (or time-frequency regions) and measuring the drop in the HFO detection score; similarly, insertion was calculated by revealing only the top-attributed region.

EHR: the faithfulness to feature deletion was calculated by removing features ranked top-k SHAP ($k \in \{1, 3, 5, 10\}$) and measuring the expected drop in predicted probability and/or Area Under the Receiver Operating Characteristic (AUROC) on the validation fold.

Multimodal emotion recognition: deletion of modality (ablation) served as a faithfulness proxy by masking one modality at a time and measuring the drop in macro-F1/accuracy, consistent with interpretable fusion analysis.

3.6.2 *Robustness (stability under perturbations)*

For each modality, explanations were recomputed under clinically non-informative perturbations and compared

using similarity metrics:

Imaging: small rotations/flips within augmentation ranges; additive noise; mild brightness/contrast.

EEG: small time-jitter; variation of notch parameter within plausible ranges; additive noise at controlled Signal-to-Noise Ratio (SNR).

EHR: bootstrap resampling within folds; alternative SMOTE random seeds; slight feature noise within measurement tolerance.

3.6.3 *Clinical plausibility (human-based)*

Plausibility was assessed using structured heuristics (e.g., anatomical alignment for CXR; alignment to time-frequency HFO bursts for EEG; clinically meaningful functional/cognitive predictors for EHR; consistency of modality contributions for emotion recognition). The absence of blinded clinician evaluation and inter-rater reliability (e.g., Cohen's Kappa) is reported as a limitation and is positioned as future work.

3.7 *Consistency of the cross-modal sampling and split rationale*

Each modality followed the standard benchmark protocol for its source dataset, resulting in different sample sizes that reflect domain constraints and dataset curation. For CheXpert, the validation partitions (n = 200) and the test partitions (n = 500) follow the widely used benchmark configuration provided with the data set and were retained to preserve comparability with the existing literature. For EEG, subject-level variability motivates reporting both aggregate and per-subject performance, and window sampling was used to balance classes and stabilize training. For CREMA-D emotion recognition, five independent random splits (70/15/15) were used to estimate variability under multimodal fusion. For predicting Alzheimer's disease in the EHR, 10-fold stratified cross-validation was used due to moderate class imbalance and the need for stable estimation under resampling (SMOTE within folds).

4. Results

Multimodal evaluation provides critical information on the predictive performance and clinical applicability of XAI in various healthcare tasks. The results are presented by modality and complemented by cross-domain synthesis. Table 2 provides a unified summary of the tasks, metrics, and principal findings.

Table 2. Summary of multimodal evaluation

Modality	Clinical task	Metric	Result	Key findings
Chest X-ray	Pathology detection	AUC (best label)	0.84	Region-based explanations showed the highest anatomical plausibility for pulmonary abnormalities
EEG	Epilepsy decision support (HFO)	Sensitivity	0.87	Temporal characteristics aligned with clinical HFO markers
Multimodal	Emotion recognition	Accuracy	89.3%	No single dominant modality across emotional states; an interpretable fusion is required
EHR	Alzheimer prediction	AUC	0.94	MMSE-related measures contributed substantially; clinically plausible attribution of characteristics

4.1 *Predictive performance across models*

The credible predictive performance of each subsystem was verified before interpreting the explanations, presented in Table 2. In the medical imaging system, a DenseNet-121 model trained on CheXpert demonstrated consistent

discriminative ability across representative thoracic findings, with AUROC values of 0.78 (atelectasis), 0.88 (cardiomegaly), 0.82 (consolidation), 0.85 (edema), and 0.90 (pleural effusion), as shown in Figure 9. These values indicate that the model learned a non-trivial radiographic signal and supports a meaningful explanation, as presented in Figure 10.

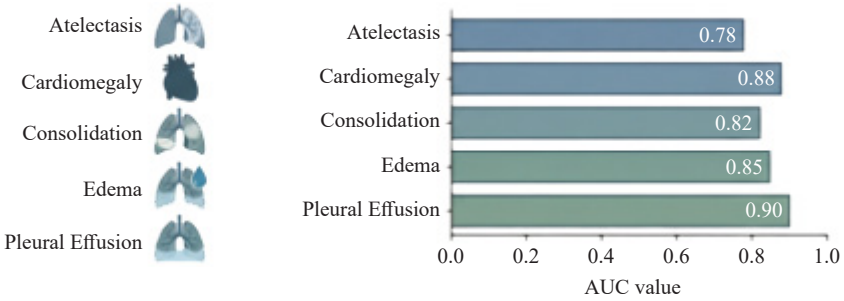


Figure 9. Illustrative AUC values from the test-set for DenseNet-121 on representative CheXpert pathologies (e.g., atelectasis, cardiomegaly, consolidation, edema, pleural effusion), providing a performance baseline for the explanation analysis

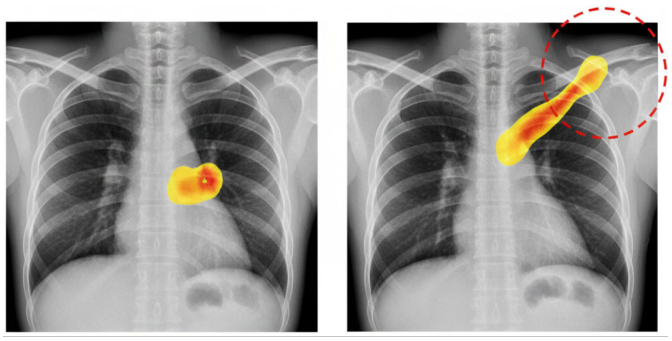


Figure 10. Region-based explanation behavior of XRAI in chest X-ray interpretation. The method generally produces cleaner, more coherent saliency regions. However, it may merge adjacent anatomical structures into a single region (e.g., the rib outline and an adjacent nodule), reducing clinical interpretability

In the structured EHR system, Alzheimer's disease prediction used SMOTE with a 10-fold stratified cross-validation and selected CatBoost as the final model. The tuned CatBoost achieved an accuracy of 95.0%, macro-F1 of 0.93, and AUROC of 0.9372. The confusion matrix indicated 12 false negatives and 10 false positives, which is clinically relevant because false negatives represent missed cases that could delay intervention.

The EEG task explored HFO-guided epilepsy decision support, focusing on interpretable evidence grounded in clinically relevant biomarkers and the time-frequency structure. The key outcome was alignment of interpretable temporal characteristics with HFO-related clinical markers, with a sensitivity of 0.87 [7]. The multimodal mental-health task developed a trimodal system for emotion recognition, achieving performance improvements through the fusion of complementary modalities (accuracy = 89.3%) and requiring fusion-level reasoning and modality-contribution analysis [2, 12].

4.2 *Imaging explanations: qualitative comparison of XAI methods*

For identical chest radiograph cases and labels, the vanilla gradients were highly noisy and fragmented, limiting clinical usability. Grad-CAM generated smoother, more stable coarse regions that often aligned with plausible pathological locations [6]. Integrated Gradients produced more distributed attributions that can reflect spatially diffuse findings but may reduce specificity [13]. Region-based explanations were typically the most anatomically coherent. However, region aggregation occasionally merged adjacent anatomical structures. It obscured fine boundaries. These

findings reinforce the idea that explanation readability and faithfulness are not automatically aligned and motivate the inclusion of stability and faithfulness checks in the evaluation [4].

4.3 *Feature-level explanations in EHR modeling*

In the Alzheimer's EHR system, a feature-importance analysis identified Functional Assessment, ADL, Memory Complaints, MMSE, and Behavioral Problems as the dominant contributors to the model's decisions. These drivers are clinically plausible and align with diagnostic reasoning. However, feature-level explanations still require robustness checks because correlated predictors and handling of imbalanced data can affect attribution stability, consistent with concerns in the clinical risk prediction and neuro-symbolic explainability literature [5, 8].

4.4 *Cross-modal synthesis of explainability behavior*

Cross-modal synthesis reveals modality-specific expectations for explanations and failure modes. Imaging explanations are intuitive because they are spatial, but remain sensitive to gradient noise, preprocessing, and spurious cues. Tabular explanations are traceable because they map to explicit variables, but they can conceal confounding and instability. Signal-based explanations are most robust when grounded in biomarkers and time-frequency evidence, but remain susceptible to artifacts and preprocessing. Multimodal explanations must also demonstrate how diverse evidence is integrated and whether the contributions of different modalities remain consistent across various contexts. Table 3 summarizes the comparative properties of the explanation mechanisms, and Table 4 contrasts the multimodal findings with the common assumptions of a single mode.

Table 3. Comparative properties of explanation mechanisms across modalities

XAI method	Fidelity	Robustness	Clinical plausibility	Computational cost
Grad-CAM [4]	Medium	High	Medium-High	Low
Integrated gradients [16]	Medium	Medium	Medium	Medium
Region-based (XRAI-style)	High	High	High	High
SHAP-style attribution [9]	High	High	High (tabular)	Medium
Attention weights	Medium	Medium	Context-dependent	Medium

Table 4. Multimodal extensions relative to common single-modality findings

Evaluation dimension	Single-modality findings	Multimodal extensions observed here
Transferability of the method	Grad-CAM is effective for imaging	Transferability is modality-limited; modality mismatch risks increase
Prevalence of failure mode	Instability reported in imaging saliency	Instability manifests across modalities (spatial drift vs rank shifts vs preprocessing sensitivity)
Robustness requirements	Domain constraints help physiological XAI	Constraints improve interpretability reliability across modalities
Explanation universality	SHAP proposed as general-purpose	Effectiveness remains modality-dependent; strongest for tabular, weaker without biomarker grounding

5. Discussion

The consolidated evidence indicates that explainability is a modality-dependent property that emerges from interactions between data type, model architecture, and clinical context [2, 4]. In chest X-ray prediction, the quality of the explanation is constrained by gradient behavior and feature map resolution, which helps explain the differences among vanilla gradients, Grad-CAM, integrated gradients, and region-based approaches. The practical implication is that heatmap-based explanations should be treated as clinician-facing communication aids and model debugging tools, not as proof of causal reasoning, consistent with established concerns about faithfulness [2].

For EHR-based Alzheimer prediction, interpretability aligns more naturally with clinical reasoning because explanations reference named variables. However, this does not eliminate the need for careful validation; feature importance can reflect correlated proxies or the effects of imbalance correction. Consequently, patient-facing tabular explanations should report how imbalance handling was performed and provide evidence that feature rankings are stable between folds and sampling strategies, in accordance with the responsible CDSS recommendations [3, 5, 8].

To support the EEG/HFO decision, interpretability is most defensible when grounded in domain markers and time-frequency evidence, consistent with medically oriented XAI principles [7] and timing-based evidence for identifying epileptogenic zones [8]. In multimodal emotion recognition, interpretability must address fusion: explanations should justify both the decision and each modality's contribution, reflecting broader multimodal AI challenges in healthcare and mental health contexts [2, 12].

These results motivate a modality-aligned guide rather than a universal recommendation. Table 5 summarizes the practical recommendations for clinical implementation.

Table 5. Modality-specific recommendations for clinician-facing deployment

Modality	Recommended explanation approach	Rationale
X-ray	Grad-CAM [4] + region-based explanations	Balances interpretability and stability; supports clinician communication
EEG	Biomarker/time-frequency grounded interpretation [12]	Generic saliency insufficient; domain markers improve validity [7]
Multimodal	Attention analysis + modality ablation [5]	Needed to justify fusion and contribution stability
EHR	SHAP-style attribution + stability checks [9]	Variable-level explanations are traceable but must be robust

5.1 *Limitations and future directions*

This study has several limitations that should be considered when interpreting the results and deploying the proposed framework. First, the four modality pipelines use different datasets, splits, and model families, which limits direct numerical comparability and motivates our emphasis on modality-aligned evaluation principles. Second, an evaluation of clinical plausibility by clinicians was not conducted; instead, plausibility judgments were based on literature-informed rubrics, and future research will include expert panels and report inter-rater reliability. The clinical plausibility findings in this work are preliminary and need formal validation by clinicians. Third, the quality of the imaging explanation was primarily qualitative; quantitative metrics for faithfulness and stability (e.g., insertion/deletion AUC, perturbation consistency) should be computed in future work, when per-sample predictions and explanation artifacts are available. Fourth, the CheXpert validation/test sizes reflect the benchmark split, and additional external validation datasets would further strengthen generalizability. Fifth, while cloud and big data deployment considerations are discussed, the experiments were conducted on a single workstation without distributed training; future work should explicitly evaluate scalability in cloud and federated settings.

Future work will prioritize: (i) cloud-based distributed XAI evaluation for large-scale clinical data, including batching and approximate explanation strategies; (ii) federated multimodal XAI scenarios with explanation aggregation and privacy-aware logging; (iii) clinician-in-the-loop studies to quantify plausibility and workflow impact; and (iv)

controlled-variable experiments that isolate modality effects by applying a shared explainer across matched architectures where feasible [17-20].

6. Conclusions

This paper consolidated four healthcare decision support pipelines under a unified evaluation perspective of explainability spanning: (i) chest radiograph pathology classification, (ii) EEG-based epilepsy decision support using High-Frequency Oscillations (HFOs), (iii) multimodal emotion recognition for psychological decision support, and (iv) prediction of Alzheimer's disease based on EHR. Across these modalities, we examined how commonly used explanation mechanisms behave when applied to fundamentally different types of clinical evidence types; spatial patterns in imaging, temporal/time-frequency structure in biosignals, modality fusion in behavioral inference, and variable-level reasoning in tabular EHR data.

In the imaging setting, DenseNet-121 achieved meaningful discrimination of representative CheXpert findings, enabling comparative interpretation with gradient- and region-based methods. Qualitatively, Grad-CAM tended to provide stable, coarse localizations aligned with plausible anatomical regions; integrated gradients produced smoother, but sometimes more diffuse attributions; region-based explanations improved anatomical coherence while occasionally merging adjacent structures; and vanilla gradients were frequently noisy and fragmented. In the EHR setting, the tuned CatBoost model yielded strong predictive performance, and the feature-level interpretation highlighted clinically significant drivers (notably functional and cognitive measures), supporting transparency aligned with clinical reasoning. For the support of the EEG/HFO decision, interpretability was most defensible when grounded in domain-specific biomarkers and time-frequency signatures, and attention/saturation outputs were interpreted as supportive evidence, conditioned on signal-preprocessing choices. In multimodal emotion recognition, fusion improved predictive performance, but the relative contribution of modalities varied across emotional classes, motivating fusion-level interpretability via attention patterns, modality contribution analysis, and ablation-based validation.

A cross-modal synthesis indicates that explainability is not a universal add-on; rather, it is a modality-dependent property that emerges from the interaction among data type, model architecture, and clinical context. Across modalities, recurring risks include instability of explanations under perturbations and susceptibility to spurious cues, reinforcing the importance of explanation validation as a deployment-facing requirement rather than a visualization exercise.

The core objective of this study was to evaluate the behavior of the explanation in heterogeneous clinical modalities and to assess whether the explanations are sufficiently faithful, robust, and clinically plausible to support the use of CDSS in real-world. First, this study presents a consolidated multimodal evaluation encompassing spatial imaging, temporal biosignals, multimodal fusion, and tabular EHR prediction, demonstrating how explanation choices and interpretation constraints differ across clinical evidence types. Second, identify recurring cross-domain failure modes: instability, spurious cue sensitivity, and modality mismatch that persist despite differences in datasets and model families, supporting the argument that explanation validity must be treated as a system property. Third, derive actionable modality-specific guidance for clinical workflows: Grad-CAM plus region-based attribution (with stability checks) for chest radiographs, biomarker/time-frequency-based interpretation for EEG/HFO pipelines, contribution-aware fusion analysis (attention + ablation) for multimodal psychological decision support, and feature-level attribution with stability reporting under imbalance handling for EHR prediction.

Future work should prioritize standardized quantitative evaluation of faithfulness and robustness across modalities, clinician-in-the-loop plausibility assessment with inter-rater reliability, and validation on larger real-world clinical datasets and in deployment settings where data drift and resource constraints shape both model behavior and the feasibility of explanations. In general, the proposed multimodal evaluation perspective supports practical deployment by aligning explanation selection and validation with modality-specific clinical evidence, robustness requirements, and workflow constraints.

From a cloud computing and data science perspective, the framework provides a practical rubric for operationalizing XAI at scale: explanation artifacts can be computed in batches, logged with model and data versions, monitored for stability drift, and audited within Machine Learning Operations (MLOps) pipelines. These properties are essential when CDSS must generate explanations for large patient cohorts and when models are retrained frequently. By explicitly

linking modality-aligned evaluation axes (faithfulness, robustness, clinical plausibility) to scalable computation and governance, the work enhances the applicability of XAI evaluation to cloud-enabled clinical analytics and big-data deployment scenarios.

Conflict of interest

The author declares no competing financial interest.

References

[1] Elhaddad M, Hamam S. AI-driven clinical decision support systems: an ongoing pursuit of potential. *Cureus*. 2024; 16(4): e57728. Available from: https://doi.org/10.7759/cureus.57728.

[2] Xu X, Li J, Zhu Z, Zhao L, Wang H, Song C, et al. A comprehensive review on synergy of multi-modal data and AI technologies in medical diagnosis. *Bioengineering*. 2024; 11(3): 219. Available from: https://doi.org/10.3390/bioengineering11030219.

[3] Labkoff S, Oladimeji B, Kannry J, Solomonides A, Leftwich R, Koski E, et al. Toward a responsible future: recommendations for AI-enabled clinical decision support. *Journal of the American Medical Informatics Association*. 2024; 31(11): 2730-2739. Available from: https://doi.org/10.1093/jamia/ocae209.

[4] Rizzo M, Conati C, Jang D, Hu H. Evaluating the faithfulness of causality in saliency-based explanations of deep learning models for temporal colour constancy. In: Longo L, Lapuschkin S, Seifert C. (eds.) *Explainable Artificial Intelligence*. Cham: Springer Nature Switzerland; 2024. p.125-142. Available from: https://doi.org/10.1007/978-3-031-63800-8_7.

[5] Mesinovic M, Watkinson P, Zhu T. Explainable AI for clinical risk prediction: a survey of concepts, methods, and modalities. *arXiv:2308.08407*. 2023. Available from: https://doi.org/10.48550/ARXIV.2308.08407.

[6] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*. 2020; 128: 336-359. Available from: https://doi.org/10.1007/s11263-019-01228-7.

[7] Chybowski B, Klimes P, Cimbalnik J, Travnicek V, Nejedly P, Pail M, et al. Timing matters for accurate identification of the epileptogenic zone. *Clinical Neurophysiology*. 2024; 161: 1-9. Available from: https://doi.org/10.1016/j.clinph.2024.01.007.

[8] Lu Q, Li R, Sagheb E, Wen A, Wang J, Wang L, et al. Explainable diagnosis prediction through neuro-symbolic integration. *arXiv:2410.01855*. 2025. Available from: https://doi.org/10.48550/arXiv.2410.01855.

[9] Cascella M, Leoni MLG, Shariff MN, Varrassi G. Artificial intelligence-driven diagnostic processes and comprehensive multimodal models in pain medicine. *Journal of Personalized Medicine*. 2024; 14(9): 983. Available from: https://doi.org/10.3390/jpm14090983.

[10] Kumar R, Sporn K, Gowda C, Paladugu P, Alnemri A, Khanna A, et al. Integrating multi-modal predictive modeling, imaging biomarkers, and EHR-linked decision support systems in spine-focused biomedical informatics. *Preprints.org*. 2025. Available from: https://doi.org/10.20944/preprints202506.1690.v1.

[11] Jaber D, Hajj H, Maalouf F, El-Hajj W. Medically-oriented design for explainable AI for stress prediction from physiological measurements. *BMC Medical Informatics and Decision Making*. 2022; 22: 38. Available from: https://doi.org/10.1186/s12911-022-01772-2.

[12] Dang N, Tran X, Mai A, Vi CT. Unveiling mental health states through multi-modal AI integration with EEG signals: a systematic review. In: *10th International Conference on the Development of Biomedical Engineering in Vietnam*. Cham: Springer Nature Switzerland; 2025. p.523-536. Available from: https://doi.org/10.1007/978-3-031-90197-3_40.

[13] Yue Z, Wang Y, He K. IG^2: Integrated gradient on iterative gradient path for feature attribution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2024; 46(12): 10478-10492. Available from: https://doi.org/10.1109/TPAMI.2024.3388092.

[14] Mehmood A, Mehmood F, Kim J. Towards explainable deep learning in computational neuroscience: visual and clinical applications. *Mathematics*. 2025; 13(20): 3286. Available from: https://doi.org/10.3390/math13203286.

[15] Zhai L, Zhao M, Zhang J, Jamil M, Naz R, Li C. A systematic review of EEG-based biomarkers for depression, anxiety, and bipolar disorder: trends in explainable artificial intelligence (XAI). *BMC Psychiatry*. 2025; 26: 90.

Available from: https://doi.org/10.1186/s12888-025-07589-3.
[16] Mungoli N. Scalable, distributed AI frameworks: leveraging cloud computing for enhanced deep learning performance and efficiency. *arXiv:2304.13738*. 2023. Available from: https://doi.org/10.48550/arXiv.2304.13738.
[17] Bogdanova A, Imakura A, Sakurai T. DC-SHAP method for consistent explainability in privacy-preserving distributed machine learning. *Human-Centric Intelligent Systems*. 2023; 3: 197-210. Available from: https://doi.org/10.1007/s44230-023-00032-4.
[18] Le Page L, Dionysio C, Boehm M. Scalable computation of shapley additive explanations. In: *Datenbanksysteme für Business, Technologie und Web (BTW 2025)*. Gesellschaft für Informatik, Bonn; 2025. p.355-376. Available from: https://doi.org/10.18420/BTW2025-16.
[19] Ducange P, Marcelloni F, Renda A, Ruffini F. Federated learning of XAI models in healthcare: a case study on parkinson's disease. *Cognitive Computation*. 2024; 16: 3051-3076. Available from: https://doi.org/10.1007/s12559-024-10332-x.
[20] Tunduny T, Shibwabo B. Explainable AI approaches in federated learning: systematic review. *JMIR AI*. 2026; 5: e69985. Available from: https://doi.org/10.2196/69985.
[21] Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019; 33(1): 590-597. Available from: https://doi.org/10.1609/aaai.v33i01.3301590.
[22] Barborica A, Mihai F, Tofan L, Oane I, Mindruta I. *Dataset of intracranial EEG during cortical stimulation evoking visual effects*. Openneuro. 2024. Available from: https://openneuro.org/datasets/ds005169/versions/1.0.0 [Accessed 22nd February 2026].
[23] Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R. CREMA-D: crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*. 2014; 5(4): 377-390. Available from: https://doi.org/10.1109/TAFFC.2014.2336244.
[24] El Kharoua R. *Alzheimer's disease dataset*. Kaggle. Available from: https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset [Accessed 22nd February 2026].

Appendix A

A.1 *Experimental environment and reproducibility*

All experiments were executed on a single compute node (no distributed training was used). The experimental environment is summarized in Table A1. For deep learning pipelines, training and inference were implemented in Python using a modern deep learning framework (e.g., PyTorch) with GPU acceleration where available; classical ML pipelines were implemented using scikit-learn, with CatBoost used for gradient-boosted decision trees. For explainability, gradient-based methods were implemented using standard attribution libraries (e.g., Captum for PyTorch), and SHAP-style explanations for tree models were computed using TreeSHAP implementations in SHAP/appropriate libraries.

Table A1. Experimental environment and software versions

Component	Specification
Central Processing Unit (CPU)	16-32 logical cores (Xeon/Ryzen class), AVX2 capable
Graphic Processing Unit (GPU)	NVIDIA RTX 3090 (24 GB)
Random-Access Memory (RAM)	64 GB
Operating System (OS)	Ubuntu 22.04 LTS
Python	Python 3.11.x (compatible with SHAP ≥ 0.50)
Deep learning framework	PyTorch 2.11 (CUDA-enabled)
Key libraries	Numpy ≥ 1.26; pandas ≥ 2.2; scipy ≥ 1.12; scikit-learn 1.7.x-1.8.x; imbalanced-learn ≥ 0.12; matplotlib ≥ 3.8
Model/XAI libraries	Catboost 1.2.10; shap 0.51.0; captum 0.8.0; opencv-python 4.13.0.92; mediapipe 0.10.33; transformers 5.4.0 (if used)

A.2 *Model hyperparameter and training configurations*

To support reproducibility and cross-modal comparison, Table A2 reports the complete hyperparameter configuration for each model: DenseNet-121 (CXR), HFONet1D and HFOTransformer1D (EEG), E-Branchformer fusion model (emotion), and CatBoost (EHR Alzheimer prediction).

Table A2. Hyperparameter configuration details of models

Model (Modality)	Key architecture settings	Optimizer	LR/Schedule	Batch/ Epochs	Regularization/ Stability	Split protocol/ Notes
DenseNet-121 (CXR; CheXpert)	121-layer DenseNet; 14 sigmoid heads; input 320 × 320	Adam	1e-4/constant	64/5 epochs	$\beta_1 = 0.9$, $\beta_2 = 0.999$; checkpoint every 4,800 iters	Uncertainty handling U-ones; official split (~ 223 k train/ 200 val/500 test)
HFONet1D (EEG; iEEG HFO)	1D CNN for 0.2 s windows; HFO vs non-HFO	Adam	1e-3/ ReduceLROnPlateau	256/10-30 epochs	Dropout 0.2; grad clip 1.0	Subject-wise split
HFOTransformer1D (EEG; iEEG HFO)	1D transformer; attention visualized from early layer	AdamW	3e-4 + cosine decay + warmup 5%	128/20-50 epochs	Dropout 0.1; wd 0.01	Subject-wise split

Table A2. (cont.)

Model (Modality)	Key architecture settings	Optimizer	LR/Schedule	Batch/ Epochs	Regularization/ Stability	Split protocol/ Notes
E-Branchformer trimodal fusion (Emotion; CREMA-D features)	Per-modality encoders hidden 2,048, 8 heads, latent 50; shared fusion 2 layers; shared rep 512; classifier 512 → 1,536 → 512 → 6	Adam	LR 5e-5, wd 5e-6, β = (0.95, 0.999); warmup 500 steps then linear decay	Batch 8; up to 50 epochs; early stopping by macro-F1	Dropout 0.25; grad-norm clip 0.3; grad accumulation 2; mixed precision	Five splits (70/15/15), fixed seeds (42 + k); no overlap checks
CatBoost (EHR; Alzheimer's)	CatBoost selected from 8 models; 10-fold stratified CV; SMOTE in the pipeline	CatBoost built-in	Learning_rate 0.05	Iterations 1,000; depth 6; l2_leaf_reg 3	-	Macro-F1 selection; AUROC reported; feature importance for interpretation

A.3 *Implementation details of XAI methods*

To improve reproducibility and address modality-specific configurations, Table A3 reports implementation settings for each XAI method, including target layer selection for Grad-CAM, baseline/steps for Integrated Gradients, SHAP variant, and background strategy for tabular attribution, and attention normalization used for fusion interpretability.

Table A3. Implementation settings for explainability methods

Methods	Modality	Implementation	Key parameters	Output artifact
Grad-CAM	CXR	Gradients w.r.t. last convolutional block of DenseNet-121; target = pathology logit (pre-sigmoid)	Target layer: last conv feature map; upsample = binary; normalize [0, 1] per image	Heatmap overlay; optional quantitative insertion/ deletion curves
Integrated gradients	CXR	Path integral from baseline to input over gradients of target logit	Baseline = all-zeros; steps = 50; attribution normalized to unit-sum	Attribution map; insertion/ deletion scores
Vanilla gradients	CXR	Input-output derivative of target logit	Absolute gradients: optional smoothing (not used unless specified)	Pixel-level saliency map
XRAI-style regions	CXR	Superpixel segmentation + region ranking using IG aggregated per region	SLIC ~ 300 regions; region aggregation threshold (based on coverage)	Region masks; rank region overlays
TreeSHAP (TreeExplainer)	EHR	Exact SHAP values for the tree ensemble (CatBoost)	Background = random subset of training data (e.g., 512 rows); output scale = probability or logarithmic odds	Global feature importance; local explanations
Attention interpretation	EEG/ Emotion	Use post-softmax attention weights; summarize per head and layer	Row-wise softmax; average over heads; optionally attention rollout	Attention maps; modality contribution analysis
Saliency (gradient)	EEG	Gradient of output logit w.r.t. input time samples	Normalize to unit norm; average profiles for correct predictions	Saliency profiles over time (mean ± SD)