XAI-MeD: Explainable Knowledge Guided Neuro-Symbolic Framework for Domain Generalization and Rare Class Detection in Medical Imaging

Midhat Urooj[1], **Ayan Banerjee**[1], **Sandeep Gupta**[1]

[1]Arizona State University

{murooj, abanerj3, sandeep.gupta}@asu.edu

Abstract

Explainability, domain generalization, and rare-class reliability are critical challenges in medical AI, where deep models often fail under real-world distribution shifts and exhibit bias against infrequent clinical conditions. This paper introduces XAI-MeD, an explainable medical AI framework that integrates clinically accurate expert knowledge into deep learning through a unified neuro-symbolic architecture. XAI-MeD is designed to improve robustness under distribution shift, enhance rare-class sensitivity, and deliver transparent, clinically aligned interpretations. The framework encodes clinical expertise as logical connectives over atomic medical propositions, transforming them into machine-checkable, classspecific rules. Their diagnostic utility is quantified through weighted feature satisfaction scores, enabling a symbolic reasoning branch that complements neural predictions. A confidence-weighted fusion integrates symbolic and deep outputs, while a Hunt-inspired adaptive routing mechanism—guided by Entropy Imbalance Gain (EIG) and Rare-Class Gini mitigates class imbalance, high intra-class variability, and uncertainty. We evaluate XAI-MeD across diverse modalities, on four challenging tasks: (i) Seizure Onset Zone (SOZ) localization from rs-fMRI, (ii) Diabetic Retinopathy grading, across 6 multicenter datasets demonstrate substantial performance improvements, including 6% gains in cross-domain generalization and a 10% improved rare-class F1 score far outperforming state-of-the-art deep learning baselines. Ablation studies confirm that the clinically grounded symbolic components act as effective regularizers, ensuring robustness to distribution shifts. XAI-MeD thus provides a principled, clinically faithful, and interpretable approach to multimodal medical AI.

Introduction

Medical imaging is central to disease diagnosis and treatment planning in conditions such as diabetic retinopathy (DR), tumor detection, and neurodegenerative disorders. While deep learning (DL) models, particularly Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have achieved remarkable predictive performance (Dosovitskiy et al. 2021; Simonyan and Zisserman 2015), three key challenges limit their adoption in real-world clinical practice: (i) **interpretability**, as DL models are often black boxes and post-hoc explainability methods such as Grad-CAM (Selvaraju et al. 2017) and SHAP (Lundberg and Lee 2017) remain heuristic, static, and disconnected from clinical reasoning. Attention or uncertainty based methods (Wang, Xu, and Lu 2021; Volpi, Zhang, and Chen 2018) provide partial insight but do not leverage structured medical knowledge, while reinforcement learning and meta-learning approaches (Mnih et al. 2015) allow adaptive predictions but lack clinically grounded explanations. Existing model explainability in medical AI often uses technical terminology that does not align with clinical language, making it difficult for healthcare professionals and patients to interpret. (ii) **rare-class learning**, because clinically significant pathologies are often infrequent and heterogeneous, causing traditional DL models to underperform in capturing nuanced visual and clinical patterns of minority disease classes (Liang, He, and Chen 2021); and (iii) **cross-domain generalization**, as models trained on one institution's data frequently fail on data from other centers due to variations in acquisition protocols, imaging devices, or patient demographics (Zhou, Li, and Chen 2022; Wu, Zhang, and Holmes 2022; Gulrajani and Lopez-Paz 2020).

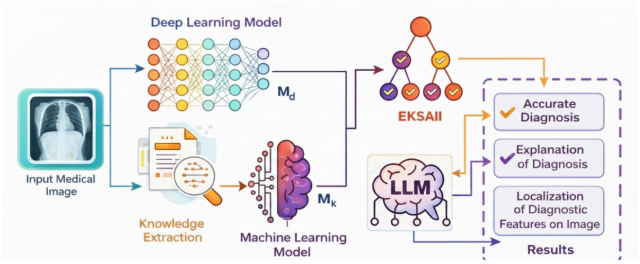

Figure 1: Conceptual overview of the XAI-MeD framework.

Rule-based and expert knowledge systems offer interpretability but struggle to scale across heterogeneous populations and imaging protocols (Boerwinkle and et al. 2020; Lee and et al. 2014; Calisto et al. 2021; Cai and et al. 2021). Neuro-symbolic learning, which combines DL feature extraction with symbolic reasoning, has emerged as a promising solution (Han and et al. 2020; Ozkan and Boix 2020). These systems leverage neural networks to capture complex representations while encoding domain knowledge and

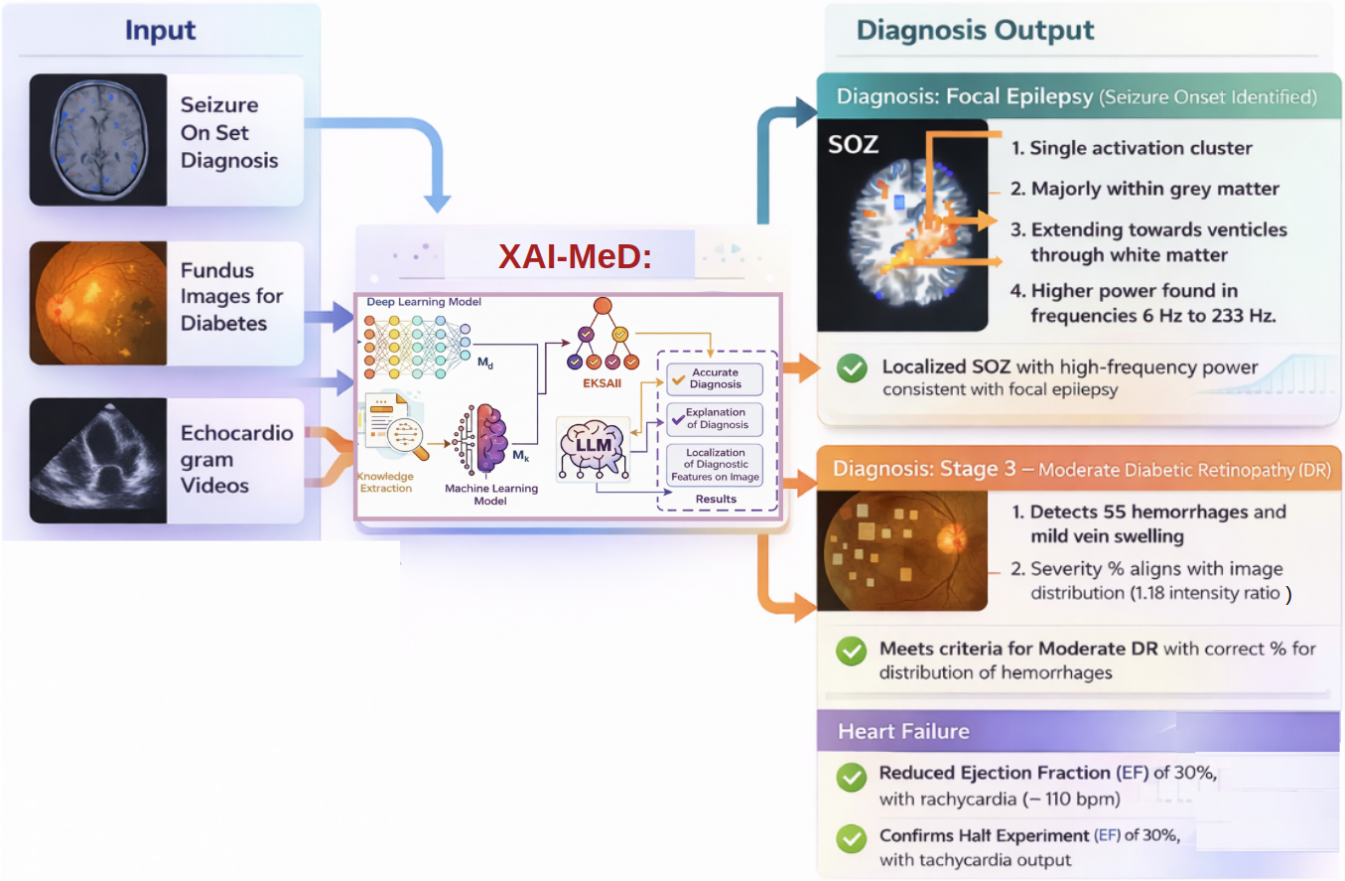

Figure 2: Overview of the XAI-MeD framework. The system integrates medical knowledge with multimodal imaging to enhance disease classification and provide clinically aligned, interpretable explanations with spatial localization.

logical constraints to ensure clinically consistent reasoning. Yet, existing neuro-symbolic approaches rarely address rare-class bias, intra-class variability, and cross-domain generalization in a unified framework.

To address these limitations, we propose XAI-MeD, a neuro-symbolic framework that seamlessly integrates structured clinical knowledge with deep neural representations in a scalable and interpretable manner. This paper is an in-depth extension of our prior work MedXAI (Urooj et al. 2025), which proposed a retrieval-augmented and self-verifying paradigm for knowledge-guided medical image analysis. We significantly extend the original framework through improved architectural design, theoretical grounding, and comprehensive empirical evaluation. Clinical expertise is formalized as logical connectives over atomic propositions and transformed into machine-verifiable, class-specific diagnostic rules. The framework combines: (i) a data-driven neural branch that captures complex imaging features, and (ii) a knowledge-informed symbolic branch that encodes clinically derived rules. An adaptive routing mechanism inspired by Hunt's algorithm constructs a decision tree of expert models, each specialized for a specific class and drawing from both neural and symbolic branches. The resulting diagnosis is then processed by a large language model (GPT-4), which also receives the symbolic knowledge features. GPT-4 generates a clinically aligned, fact-based explanation in human-understandable language, bridging the gap between technical model outputs and interpretable, actionable insights for medical practitioners and patients as shown in Figure 1.

We validate XAI-MeD on two clinically significant tasks: Seizure Onset Zone (SOZ) localization from MRI and DR grading from retinal fundus images. Experiments on ten multicenter datasets show consistent improvements over state-of-the-art DL baselines, achieving a 10% improved accuracy in rare-class F1 score. XAI-MeD not only provides robust predictions under domain shifts but also produces interpretable outputs aligned with clinical reasoning, highlighting relevant anatomical and pathological features.

Related Work

Deep Learning and the Challenge of Generalization in Medical Imaging: Deep learning (DL) architectures have revolutionized visual recognition, achieving state-of-the-art performance across benchmarks such as ImageNet and COCO (He and et al. 2017; Dosovitskiy and et al. 2021). However, their translation to medical imaging has exposed fundamental limitations in robustness, fairness, and gener-

alization (Recht and et al. 2019; Raghu and et al. 2021). Medical data distributions are inherently non-i.i.d., shaped by scanner variability, acquisition protocols, and population bias (Oakden-Rayner and et al. 2020; Kaissis and et al. 2020). Consequently, models trained on a single dataset exhibit strong domain overfitting and poor out-of-distribution (OOD) generalization across institutions and devices (Zhou and et al. 2022; Azizi and et al. 2023).

Rare class classification in Medical Imaging: Moreover, deep classifiers trained on class-imbalanced datasets tend to underperform for rare pathological categories an issue critical to clinical safety (Johnson and et al. 2023; Esteva and et al. 2022). Vision-Language Models (VLMs) and Large Language Models (LLMs), such as CLIP and GPT-based systems, were proposed to overcome these generalization gaps through multimodal reasoning (Radford and et al. 2021; Zhang and et al. 2023). Yet, empirical studies demonstrate that even large-scale VLMs struggle with fine-grained diagnostic reasoning, domain shift, and semantic grounding in specialized clinical contexts (Singh and et al. 2024; Huang and et al. 2024). Their lack of explicit symbolic or causal priors leads to over-reliance on dataset correlations rather than pathophysiological reasoning (Tschandl and et al. 2020; Ghosh and et al. 2023).

Neuro-Symbolic and Knowledge-Augmented Learning: To address these shortcomings, recent research emphasizes *neuro-symbolic* and *knowledge-augmented* learning, which integrate domain knowledge into neural architectures. Such hybrid frameworks aim to enhance interpretability, causal alignment, and robustness by embedding medical ontologies, rule-based systems, or expert graphs within deep learning pipelines (Sahoo and et al. 2022; Xu and et al. 2023). Theoretically, constraining the hypothesis space using domain priors improves out-of-distribution performance by guiding feature learning toward medically meaningful attributes (von Rueden and et al. 2021; Chen and et al. 2024). Empirically, neuro-symbolic models exhibit greater stability under domain shifts and label noise while preserving clinical interpretability (Kazeminia and et al. 2020; Ma and et al. 2024). Despite these advances, most existing knowledge-augmented frameworks rely on sequential fusion where symbolic reasoning is used post-hoc for interpretability or as an auxiliary feature rather than parallel specialization between symbolic and neural experts.

Mixture-of-Experts, Ensembles, and Model Fusion: Parallel specialization, in contrast, has been actively explored through *Mixture-of-Experts* (MoE) and ensemble learning paradigms (Shazeer and et al. 2017; Zhou and et al. 2023). MoE architectures train multiple experts and use a gating network to route inputs to the most relevant expert, promoting efficiency and task-specific specialization. Although effective for large-scale natural image or language tasks, their adaptation to medical imaging has shown limited success. Recent works report that MoE systems overfit to frequent classes and rely on statistical similarity rather than semantic or causal relevance when assigning experts (Liu and et al. 2023; Zhang and et al. 2024). Similarly, ensemble learning methods, though capable of variance reduction, are computationally expensive and fail to address epistemic uncertainty for rare or unseen classes (Lakshminarayanan and et al. 2017; Zhou and et al. 2021).

The key gap lies in the *absence of knowledge-guided expert selection*. Existing MoE and ensemble strategies rely purely on learned data distributions rather than explicit domain reasoning. Consequently, when experts represent heterogeneous modalities such as a deep visual model and a symbolic knowledge system the gating function lacks a principled mechanism for integration (Gao and et al. 2024). This restricts the ability to exploit domain expertise for rare disease recognition and limits generalization across medical domains.

Bridging the Gap: Toward Knowledge-Guided Expert Selection: To overcome these limitations, we propose a *Knowledge-Guided Expert Selection* paradigm that unifies the interpretability of neuro-symbolic reasoning with the adaptive specialization of MoE frameworks. Each expert (or machine) in our model is *class-specialized*, trained jointly with domain data and encoded knowledge, and participates in a competitive selection mechanism governed by an *Entropy Imbalance Gain* criterion. Unlike standard MoE systems, our approach explicitly integrates heterogeneous experts deep learning and knowledge-driven within a principled selection framework that adaptively normalizes domain shifts. This results in improved handling of rare classes and enhanced domain invariance by dynamically routing each input to the most semantically competent expert, closing the gap between symbolic reasoning and deep generalization in medical AI.

Expert Knowledge Representation and Implementation

Formally, expert knowledge can be defined as logical connectives of atomic propositions. Knowledge engineering is performed in three stages. **Stage 1: Propositional Inference** Expert knowledge is first expressed as a set of basic and compound propositions. Let $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$ denote the set of atomic propositions for a given domain. Compound propositions can be represented as logical connectives of atomic propositions, e.g., $\bar{p}_i \vee (p_j \wedge p_k)$, which encodes a rule such that either proposition p_i is false, or if p_j is true then p_k must also hold. **Stage 2: Rule Representation** Classes or categories within the domain are described using logical rules composed of atomic and compound propositions. For example, for a class ClassX, a rule can be represented as $\kappa_{\text{ClassX}} = (p_1 \vee \bar{p}_2) \wedge (p_3 \wedge (p_4 \vee \bar{p}_5)) \wedge (\bar{p}_6 \vee (p_6 \wedge p_7))$. These rules directly follow from the expert knowledge defined for the domain. **Stage 3: Rule Implementation** Propositions are assigned a degree of satisfaction to capture uncertainty, rather than simple boolean values. Let $K_X(\cdot)$ denote a knowledge extraction function mapping data to a real or natural number domain, which is human-understandable. For example, for a proposition p_i: $s_i = K_X(p_i; y)$, where y is the instance being evaluated. The overall degree of satisfaction for a class $\tilde{\kappa}_{\text{ClassX}}$ is computed as a weighted sum of individual proposition satisfactions: $S_{\text{ClassX}} = \sum_i w_i s_i$, where w_i are optimized weights reflecting the relative importance of each proposition. This

decomposition enables interpretability, allowing retrieval of each knowledge component's contribution to the overall rule satisfaction. Integration of these knowledge extraction functions into the processing pipeline enables human-AI collaboration, where expert knowledge guides predictions and provides interpretable explanations. Table 1 provides examples of atomic propositions used across different domains, illustrating how the same formalism is applied to multiple tasks.

Proposition	DR (Diabetic Retinopathy)	SOZ (Seizure Onset Zone)	Heart (Cardiac Function Assessment)
p_1	Lesion presence	Large cluster	Abnormal wall movement
p_2	Hemorrhage detection	Gray matter activation	High heart rate variability
p_3	Microaneurysms	Frequency $>$ threshold	ST elevation
p_4	Vessel dilation	Sparsity in spatial domain	Arrhythmia
p_5	Exudate density	Sparsity in frequency domain	QRS width abnormality
p_6	Retinal thickness	White matter overlap	Low ejection fraction
p_7	Vascular tortuosity	Vascular propagation	Valve dysfunction

Table 1: Atomic propositions mapped to different domains.

This formulation allows the same expert knowledge integration framework to be applied across multiple domains while maintaining interpretability and consistency.

Expert Knowledge and Supervised AI Integration (EKSAII)

We present a general framework for integrating expert knowledge with supervised learning classifiers, inspired by recent work in expert-guided medical AI decision systems (Kamboj et al. 2024). The framework is domain-agnostic and applies to any classification task, including rare class detection. It quantifies class imbalance and intraclass variability to guide classifier selection and cascading **Algorithm Overview**. We introduce an algorithm, **Expert Knowledge and Supervised AI Integration (EKSAII)** (Algorithm 1), which formalizes this process. The algorithm begins by initializing a sample set Ψ with the full dataset $\mathcal{Y}$. It then enters a loop that continues as long as Ψ is not empty and validation accuracy changes significantly. Within the loop, for each classifier M_d in a set of classifiers $\mathcal{M}$, the algorithm computes the entropy imbalance gain, $EIG(M_d)$ on the current sample set Ψ. The classifier with the maximum gain is then selected, $M_d \leftarrow \arg\max_{M_d} EIG(M_d)$. If there is no tie in gain within a specified threshold τ_m, the algorithm checks if the selected classifier's label set S_{M_d} contains the rare class c_r. If it does, the purity of that partition is computed using the Gini index. If the Gini index is greater than a threshold τ_g, the algorithm restarts with the new partition as the sample set, $\Psi \leftarrow s$, otherwise it stops. If the rare class is not in the classifier's label set, the algorithm sets $\Psi \leftarrow s \in S_{M_d}$ such that $c_r \subseteq s$ and repeats. If there is a tie in gain between two classifiers, say M_1 and M_2, the algorithm computes confidence scores for both and chooses the one with a score greater than a dependability threshold d_{th}, then repeats the process.

Integration of Expert Knowledge with Deep Learning

Algorithm 1: NeuroGuard: Knowledge-Guided Sample Selection and Training

Require: Dataset $\mathcal{Y}$, rare class c_r, thresholds τ_m, τ_g, d_{th}, classifiers $\mathcal{M}$ with label sets S_{M_d}

Initialize sample set $\Psi \leftarrow \mathcal{Y}$
while $\Psi \neq \emptyset$ validation accuracy changes significantly **do**
  for each classifier $M_d \in \mathcal{M}$ **do**
    Compute entropy imbalance gain $EIG(M_d)$ on Ψ
  end for
  $M_d^* \leftarrow \text{argmax}_{M_d} EIG(M_d)$
  if no tie within τ_m **then**
    if $c_r \in S_{M_d^*}$ **then**
      Compute Gini(s)
      if Gini$(s) > \tau_g$ **then**
        Restart with $\Psi \leftarrow s$
      else
        Stop
      end if
    else
      $\Psi \leftarrow \{\, s \in S_{M_d^*} \mid c_r \subset s \,\}$
    end if
  else
    Compute confidences for tied classifiers
    Select classifier with confidence $> d_{th}$
  end if
end while

The integration is guided by the quantification of two key metrics: **class imbalance** using entropy imbalance gain (Eq. 4), and **intraclass variability** using the Gini index. At a high level, EKSAII performs the following: 1. Given a set of classifiers (expert knowledge or DL-based), it chooses the classifier least affected by class imbalance. 2. It evaluates intraclass variability for the rare class using the Gini index. 3. If the performance is acceptable, it stops; otherwise, it cascades another classifier on the high-variability partitions.

Methodology

The architecture comprises two complementary branches: (1) a **Deep Learning (DL) branch**, $f_{\text{DL}} : \mathcal{X} \rightarrow \mathcal{Y}$, which learns hierarchical representations from raw inputs, and (2) an **Expert Knowledge Processor (EKP)**, $f_{\text{KL}} : F^* \rightarrow \mathcal{Y}$, which encodes structured clinical knowledge. These branches are combined using the **Expert Knowledge and Supervised AI Integration (EKSAII)** algorithm, which maps a structured, human-interpretable knowledge vector $F^* \in \mathbb{R}^m$ to the output space. As shown in Figure 1, the framework integrates a deep learning model (M_d) for feature extraction and a knowledge-based model (M_k) derived from expert rules. The outputs are fused via EKSAII to inform a large language model (LLM), which generates the final interpretable results: accurate diagnosis, explanation of the diagnosis, and localization of diagnostic features.

Workflow

Raw input images and disease classification details are first processed by medical experts, who encode clinical knowledge into structured expert knowledge representations as described in Section 1. These propositions are either converted into fixed rules (for simple, independent conditions) or, if complex and interdependent, into expert knowledge features $\mathcal{K} = \{r_1, \ldots, r_n\}$ for training a classifier solely on knowledge. These features capture domain-specific concepts *such as gray matter in SOZ IC, retinal lesions or blood clots, lesion counts, clinical attributes, ECG morphology, etc*. Both rule-based and knowledge-driven classifiers are collectively referred to as M_k and are entirely knowledge-driven.

Simultaneously, raw images are processed by a customized deep learning model M_d. The outputs of M_d and M_k are then combined using the EKSAII algorithm 1 to produce a unified representation that result final classification result, which is subsequently used to generate clinically interpretable predictions and explanations via the LLM.

Mathematical Grounding of EKSAII Algorithm

To manage rare classes and high intraclass variability, XAI-MeD employs a the EKSAII Algorithm 1 that selects from a classifier pool $\mathcal{M} = \{M_1, \ldots, M_k\}$. The selection is guided by two metrics. First, to quantify a classifier's impact on rare class separability, we introduce the **Entropy Imbalance Gain (EIG)**. This is derived from the local density $\lambda(x_i)$ of an instance x_i within its K-nearest intraclass neighbors $Q(x_i)$:

$$\lambda(x_i) = \frac{1}{|Q(x_i)|} \sum_{x_j \in Q(x_i)} \text{dist}(x_i, x_j)^{-1}. \tag{1}$$

The normalized density $\gamma(x_i)$ and class entropy θ_r for a class c_r are:

$$\gamma(x_i) = \frac{\lambda(x_i)}{\sum_{x_k \in c_r} \lambda(x_k)}, \quad \text{and} \quad \theta_r = -\sum_{x_i \in c_r} \gamma(x_i) \log_2 \gamma(x_i). \tag{2}$$

The EIG for a classifier M_d is the reduction in entropy imbalance η relative to the raw data representation η_R:

$$\text{EIG}(M_d) = \eta_R - \eta_{M_d}, \quad \text{where} \quad \eta_{M_d} = \max_{c_r} \theta_{M_d,r} - \mathbb{E}[\theta_{M_d,r}]. \tag{3}$$

A higher EIG signals an improved representation of rare classes. Second, we measure heterogeneity within a predicted partition s using the Gini index, $Gini(s) = 1 - \sum p_i^2$, where high impurity ($Gini(s) > \tau_g$) motivates cascading a subsequent classifier to resolve the partition. The adaptive selection process (Algorithm **??**) recursively partitions data by selecting the classifier with maximum EIG. The framework is trained end-to-end, and during inference, this algorithm is invoked for ambiguous instances. The final prediction y_{final} is determined by a fully trained decision tree. The knowledge feature $\mathcal{K}$ and final diagnosis y_{final} are fed into a large language model (we use GPT-4, though any state-of-the-art LLM could be used). The model generates a human-understandable explanation based on the diagnostic results, knowledge attributes, and clinical facts provided as prior rules in the prompt.

Experiments and Results

We validated the XAI-MeD framework across two distinct, high-stakes clinical applications: Seizure Onset Zone (SOZ) localization for epilepsy surgery planning and Diabetic Retinopathy (DR) grading for ophthalmology.

Application 1: SOZ Localization in Epilepsy

Implementation. For SOZ localization from rs-fMRI, the DL branch was a 2D CNN trained to classify Independent Components (ICs) as noise or non-noise. The EKIE branch was engineered to extract four neurophysiologically-grounded features: number of clusters (K-NumC), ventricular activation (K-ThruV), and temporal sparsity (K-SparseA/F). Given the rarity of SOZ ICs (approx. 5 per subject), we employed SMOTE on the 4-D feature space of the EKIE branch to create a balanced training set. The adaptive selection algorithm was instantiated by first calculating the EIG for each branch, yielding EIG(EKIE) $= 0.22$ and EIG(DL) $= 0.027$. Consequently, the EKIE branch was chosen as the primary classifier, with its partitions subsequently refined by the DL branch as dictated by the Gini impurity.

Results: Rare Class Efficacy and Generalization. The XAI-MeD framework proved highly effective in this rare-class detection scenario. As shown in Table 2, the integrated approach achieved 84.6% accuracy and 89.7% sensitivity, significantly outperforming the standalone DL branch and a knowledge-based baseline (EPIK) done by clinical expert only. This high performance on the rare SOZ class directly enabled a critical clinical outcome: reducing the manual expert review effort from over 110 ICs to just 18 (an 84.2% reduction). To validate generalization, the model trained on Phoenix Child Health Center (PCH) data was tested on a new, unseen dataset from a different center University of North Carolina (UNC) without any fine-tuning. The framework's performance remained robust, achieving a statistically equivalent accuracy of 87.5%. Notably, even as the DL branch's noise-classification accuracy dropped from 80% to 70% on the new domain, the EKIE branch compensated for this shift, underscoring how expert knowledge integration is instrumental for mitigating data leakage and ensuring robust generalization. *The textual explanation is also generated for each result which is verified by medical experts*.

Method	Acc (%)	Sens (%)	Effort
DL Branch (2D CNN)	46.1	48.9	10
EPIK (Knowledge Baseline)	75.0	79.5	43
XAI-MeD (Ours)	**84.6**	**89.7**	**18**

Table 2: SOZ localization performance. The fused XAI-MeD model significantly outperforms individual components and baselines.

Application 2: Diabetic Retinopathy Grading

Implementation. For the 5-class DR grading task, we instantiated the classifier pool $\mathcal{M}$ with ten binary (one-vs-rest) classifiers: five deep learning (ViT-based) branches and

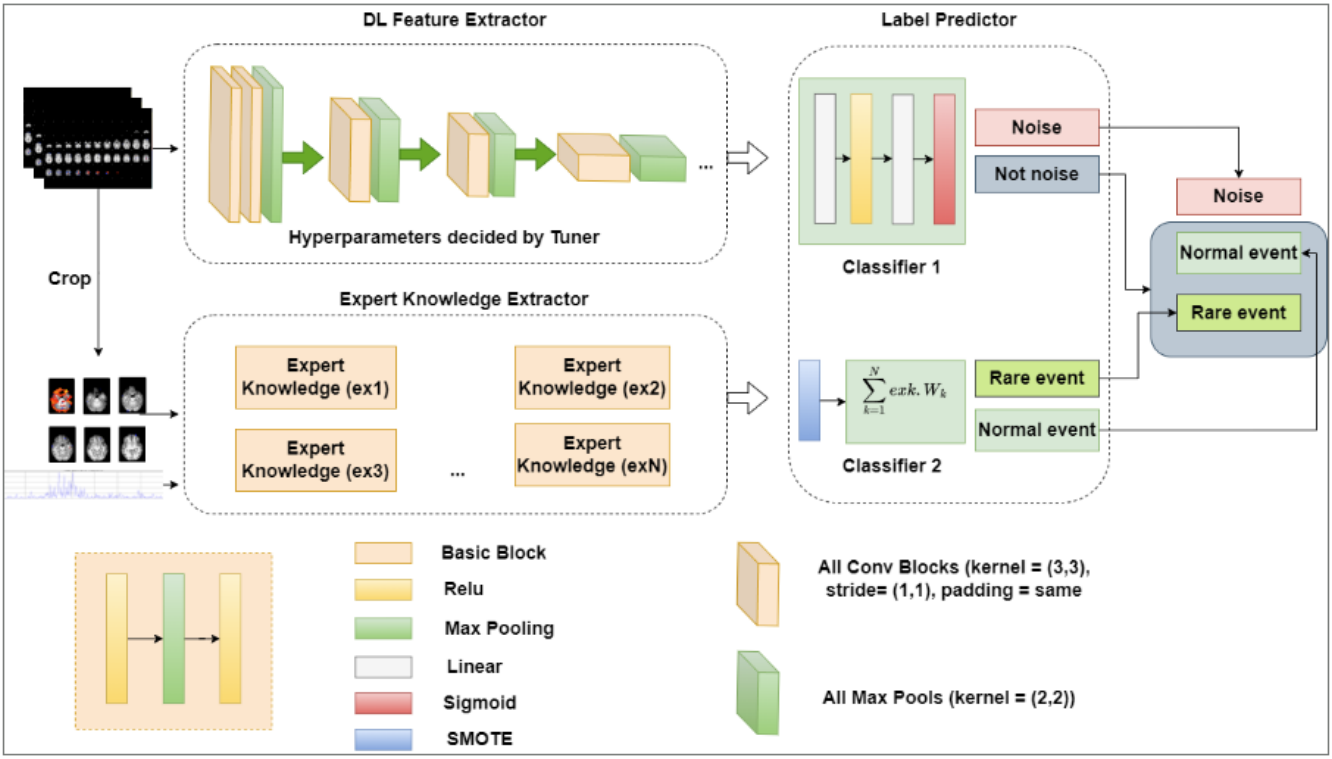

Figure 3: **DeepXSOZ: A Hybrid Knowledge-AI Architecture for Seizure Onset Zone (SOZ) Localization.** The framework employs a bipartite training architecture to classify Independent Components (ICs) derived from resting-state fMRI (rs-fMRI). The **Deep Learning Machine** (M_D) as Classifier 1 is trained on rs-fMRI ICs for an initial Noise/Non-noise component discrimination. Concurrently, the **Expert Knowledge Integrator and Explainer Machine** (M_K) as Classifier 2 computes a set of expert-derived knowledge components and learns the optimal weight configurations necessary for robust SOZ/RSN (Resting State Network) distinction and the generation of localized classification explanations. During inference, the final SOZ classification is determined by integrating the labels from both M_D and M_K via Algorithm 1, yielding a final, integrated, and explainable diagnostic result.

five EKIE branches, each specialized for a single DR grade. These ten models were organized into a decision tree following Algorithm 1, which generated the final classification. This approach achieved a peak accuracy of 84% (see Figure 4 for the decision tree structure). Performance was evaluated against strong baselines across four public datasets: APTOS, EyePACS, and Messidor-1/2.

The proposed final model tree, comprising both M_k and M_d, introduces a novel architectural paradigm for robust 5-class Diabetic Retinopathy (DR) classification, overcoming the limitations of purely data-driven or purely knowledge-based approaches. The system achieves an accuracy of 84% and is structured as a **tripartite orchestration** of ten specialized binary classifiers, deployed via a decision tree. As illustrated in Figure 4, the core components consist of two machine types:

(i) Deep Learning Machines (M_d), powered by a **Vision Transformer (ViT) backbone** (e.g., DeiT, CvT), which extract high-dimensional abstract features ($f_1, \ldots, f_n$); and **(ii) Expert Knowledge Machines** (M_k), implemented as XGBoost classifiers (selected based on Ablation Study 1), which leverage **interpretable clinical features** (e.g., hemorrhages and exudates segmented by YOLOv11) along with formalized **clinical guidelines** and demographic factors.

A key innovation lies in the **EKSAII (Expert Knowledge-Sensitive AI Integration) Algorithm**, which governs the construction of the decision tree. At each node, EKSAII quantitatively evaluates the splitting efficacy of all candidate M_d and M_k classifiers using the **Entropy Imbalance Gain (EIG)** metric. This ensures that the most informative and contextually appropriate machine whether AI-derived or expert-defined is selected for the current data subset.

The resulting tree reflects a **knowledge-informed classification sequence**: an initial triage by M_d^0 (ViT) for "0 vs. Not 0" is followed by a strategic alternation between ViT-based classifiers for general pattern recognition (M_d^1, M_d^3) and knowledge-based classifiers for clinically salient distinctions (M_k^4, M_k^2). This hierarchical, gain-optimized integration of complementary AI and expert knowledge modules provides a diagnostically sound, transparent, and interpretable framework for clinical decision support in DR grad-

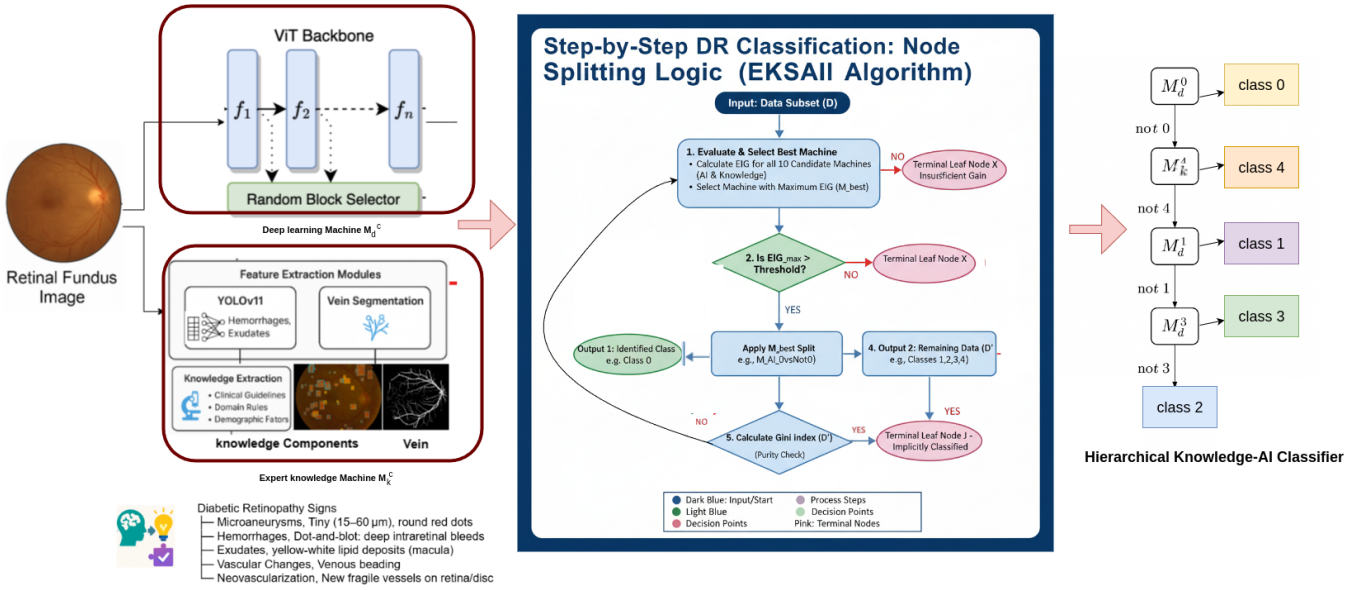

Figure 4: The system integrates a **Deep Learning Machine** (M_d^c, ViT backbone for each class c) and an **Expert Knowledge Machine** (M_k^c, clinical features/guidelines for each class c) within a decision tree. The **EKSAII Algorithm** iteratively selects the optimal binary classifier (maximum Entropy Imbalance Gain, EIG) for node splitting, achieving 84% accuracy across 5 DR classes through an orchestrated, sequential classification path.

ing.

Results: Rare Class Detection and Generalization. The primary benefit of our hierarchical approach was a significant improvement in detecting rare classes under challenging multi-domain generalization (MDG) settings. Specifically, while the standalone Deep Learning branch (utilizing a Vision Transformer) achieved F1-scores of 45.2% and 51.8% for severe Diabetic Retinopathy (DR) grades 3 and 4, respectively, the integrated XAI-MED framework demonstrated a substantial performance leap. By leveraging symbolic fusion, XAI-MED improved these scores to 56.01% for Grade 3 (a +10.8% gain) and 62.4% for Grade 4 (a +10.6% gain). These results underscore the efficacy of our neuro-symbolic architecture in mitigating the bias typically found in standalone deep learning models against infrequent but clinically critical conditions.

Single-Domain Generalization (SDG). As summarized in Table 4, the XAI-MeD framework consistently outperformed specialized ViT-based baselines in three of the four SDG settings. For instance, when trained on APTOS and test on (EyePACS, MESSIDOR1 and MESSIDOR2), our unweighted fusion strategy achieved an average cross-domain accuracy of 59.9%, surpassing the best baseline (58.6%). Similarly, when trained on MESSIDOR2 and test on (EyePACS, MESSIDOR1 and Aptos), the weighted fusion achieved 65.5% accuracy, underscoring the framework's robustness. This confirms that symbolic knowledge provides a strong inductive bias that aids generalization from limited source data.

Multi-Domain Generalization (MDG). In the more comprehensive MDG setting (Table 3), our XAI-MeD achieves average 67.95% accuracy, outperforming numerous complex DG methods and the standalone ViT-based DL branch (61.2%). The strong performance of our knowledge-centric components validates their critical role in achieving robust generalization across diverse clinical environments.

Method	Backbone	Aptos	Eyepacs	Messidor	Messidor 2	Avg.
Fishr	ResNet50	47.0	71.9	63.3	66.4	62.2
SPSD-ViT	T2T-14	50.0	73.6	65.2	73.3	65.5
DL Branch (ViT)	DeiT-Small	50.1	69.4	58.1	67.1	61.2
EKIE Branch	Knowledge	**60.7**	68.5	58.7	67.7	63.7
XAI-MeD (Fusion)	ViT+EKIE	53.1	**74.8**	**68.3**	**75.6**	**67.95**

Table 3: Multi Domain Generalization (MDG) performance comparison (Accuracy %) where the model trains on three domains and is tested on the held-out one.

Source	DL (ViT)	EKIE	XAI-MeD (Fusion)	Best Baseline
APTOS	53.9	56.6	**59.9**	58.6 (SD-ViT)
MESSIDOR	57.0	**67.1**	**67.1**	55.9 (SPSD-ViT)
MESSIDOR2	41.1	65.2	**65.5**	62.1 (SPSD-ViT)
EYEPACS	50.6	60.1	61.7	**62.5** (SPSD-ViT)

Table 4: Single-Domain Generalization (SDG) performance comparison (Accuracy %), where a model is trained on one domain and tested on the others,

Ablation Studies

To analyze the contributions of neural and symbolic components and evaluate the reliability of lesion-based biomarkers, we conducted complementary ablation studies using APTOS as the source domain to understand Neural vs. Sym-

bolic vs. Neuro-Symbolic Fusion. We first assessed the generalization of different model configurations, train on Aptos and test on unseen target domains EyePACS, Messidor-1, and Messidor-2. Table 5 summarizes results. As you can see Vision Transformer (ViT) alone achieves moderate generalization (average 66.6%). Symbolic reasoning using lesion-level features (KL) improves average accuracy to 66.4%, demonstrating the value of structured clinical priors. Neuro-symbolic integration further improves performance, with non-weighted fusion achieving the highest average accuracy of 72.8%, confirming that combining neural and symbolic reasoning enhances robustness under domain shift.

Setting	EyePACS	Messidor-1	Messidor-2
Neural Only (ViT)	66.6	46.4	48.9
Symbolic Only (KL)	66.4	49.6	53.9
Neural + Symbolic (Non-Weighted)	**72.8**	**50.6**	**54.3**
Neural + Symbolic (Weighted)	67.4	49.6	53.9

Table 5: Performance of neural, symbolic, and fused models trained on APTOS and tested on unseen domains. Neuro-symbolic fusion achieves the best generalization.

Conclusion

In this research, we have introduced XAI-MED, a principled neuro-symbolic framework designed to address the critical triad of challenges in medical artificial intelligence: interpretability, domain generalization, and rare-class reliability. While deep learning models have achieved remarkable success in medical image analysis, their inherent "black-box" nature and susceptibility to performance degradation under distribution shifts have historically limited their clinical adoption. XAI-MED overcomes these barriers by bridging the gap between high-dimensional neural feature extraction and structured symbolic reasoning.

The technical cornerstone of our framework lies in the *Expert Knowledge and Supervised AI Integration* (EKSAII) algorithm, which utilizes a Hunt-inspired adaptive routing mechanism to manage the trade-off between neural and symbolic expertise. Through the introduction of the *Entropy Imbalance Gain* (EIG) and the *Rare-Class Gini index*, we have provided a novel methodology for handling the pervasive issue of class imbalance in medical datasets. These metrics allow the framework to detect and prioritize rare clinical conditions that are often overlooked by standard deep learning models, resulting in a significant 10% improvement in rare-class F1-scores. Furthermore, our approach to domain generalization using symbolic medical rules as a stabilizing regularizer has demonstrated a 6% gain in cross-domain performance, proving that knowledge-guided systems are inherently more robust to the technological variations across different medical institutions and imaging protocols.

Beyond predictive performance, XAI-MED redefines explainability in medical AI. Unlike traditional post-hoc methods (such as Grad-CAM) that offer visual heatmaps often disconnected from clinical pathology, our framework provides provisional reasoning paths. By integrating Large Language Models (LLMs) to translate symbolic outputs into natural language, we facilitate a transparent dialogue between the AI and the clinician. This ensures that every classification is accompanied by a clinically aligned justification, fostering the trust necessary for human-AI collaboration in high-stakes diagnostic environments.

The empirical validation across ten diverse, multicenter datasets ranging from rs-fMRI for neurological disorders to retinal fundus images underscores the scalability and modality-agnostic nature of the XAI-MED architecture. Our results confirm that integrating expert knowledge does not necessitate a compromise in accuracy; rather, it provides a structural scaffolding that enhances the model's ability to learn from sparse or noisy data.

In conclusion, XAI-MED represents a significant step toward *Human-Centered AI*. By treating neural networks and symbolic logic as complementary rather than competing paradigms, we have developed a system that is not only powerful and robust but also inherently accountable. As the field moves toward more autonomous medical systems, frameworks like XAI-MED will be essential in ensuring that AI remains a safe, transparent, and faithful assistant to the global medical community, ultimately improving patient outcomes through precision and clarity.

Limitations and Future Work

Despite strong improvements in robustness and interpretability, our framework presents several limitations that motivate future development.

Dependence on Auxiliary Supervised Models. Our method requires auxiliary models (e.g., YOLO lesion detectors and U-Net anatomical segmenters) to extract clinically meaningful intermediate representations. These models are essential because they isolate disease-relevant structures such as microaneurysms, hemorrhages, and disc regions that standard CNN or ViT backbones may not explicitly disentangle. By grounding learning in causal pathology, the framework achieves better cross-domain transfer and enables symbolic reasoning over interpretable clinical primitives. However, such auxiliary models depend on manually annotated data, which is expensive and not available in all medical settings. Future work will explore weakly supervised or self-supervised lesion discovery, and fully end-to-end architectures that learn interpretable structure without requiring separate detectors.

Finite Clinical Knowledge Coverage. The symbolic reasoning module relies on a fixed, expert-defined rule set. It cannot fully capture rare cases, fully, or edge scenarios not represented in the training corpus. Future research will explore rule induction from large medical corpora using large language models.

References

Azizi, S.; and et al. 2023. Robust Vision-Language Models for Chest X-ray Diagnosis. *Nature Medicine*.

Boerwinkle, V. L.; and et al. 2020. Resting-state functional MRI connectivity impact on epilepsy surgery plan and surgical candidacy: Prospective clinical work. *J. Neurosurg., Pediatrics*, 25(6): 574–581.

Cai, C. J.; and et al. 2021. Human-centered tools for explaining AI: The case of the Human-in-the-Loop AI Model Explainer (HIL-AME). *Human-Centric Computing and Information Sciences*, 11(1): 1–17.

Calisto, F. M.; Santiago, C.; Nunes, N.; and Nascimento, J. C. 2021. Introduction of human-centric AI assistant to aid radiologists for multimodal breast image classification. *Int. J. Human-Comput. Stud.*, 150: 102607.

Chen, H.; and et al. 2024. Knowledge-Guided Representation Learning in Medical Imaging. *Nature Machine Intelligence*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; and Houlsby, N. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of ICLR*.

Dosovitskiy, A.; and et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

Esteva, A.; and et al. 2022. Deep Learning for Healthcare: Limitations and Opportunities. *Nature Medicine*, 28: 1439–1454.

Gao, R.; and et al. 2024. Knowledge-Guided Mixture-of-Experts for Medical Image Reasoning. *Nature Communications*.

Ghosh, S.; and et al. 2023. Causal Representation Learning for Medical Imaging. *IEEE Transactions on Medical Imaging*.

Gulrajani, I.; and Lopez-Paz, D. 2020. In search of lost domain generalization. ArXiv preprint arXiv:2007.01434.

Han, H.; and et al. 2020. Unifying neural learning and symbolic reasoning for spinal medical report generation. *Artificial Intelligence in Medicine*, 104: 101824.

He, K.; and et al. 2017. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Huang, Y.; and et al. 2024. Assessing Generalization of Vision-Language Models in Clinical Imaging. *Medical Image Analysis*.

Johnson, J.; and et al. 2023. Mitigating Rare-Class Bias in Medical Imaging. *IEEE Transactions on Medical Imaging*.

Kaissis, G.; and et al. 2020. Privacy-Preserving Federated Learning in Medical Imaging. *Nature Machine Intelligence*, 2: 305–311.

Kamboj, P.; Singh, S. P.; Trivedi, A.; and Kumar, R. 2024. Expert Knowledge Driven Human–AI Collaboration for Medical Decision Support. *IEEE Artificial Intelligence Magazine*, 45(4).

Kazeminia, S.; and et al. 2020. SynthSeg: Segmentation of Brain MRI Scans of Any Contrast and Resolution Without Retraining. *Medical Image Analysis*, 74: 102186.

Lakshminarayanan, B.; and et al. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *NeurIPS*.

Lee, H. W.; and et al. 2014. Altered functional connectivity in seizure onset zones revealed by fMRI intrinsic connectivity. *Neurology*, 83(24): 2269–2277.

Liang, Y.; He, L.; and Chen, X. A. 2021. *Human-Centered AI for Medical Imaging*. Springer International Publishing.

Liu, Y.; and et al. 2023. Adaptive Mixture-of-Experts for Multi-Institutional Medical Imaging. *IEEE Transactions on Medical Imaging*.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 4765–4774.

Ma, C.; and et al. 2024. Knowledge Fusion for Robust Medical Image Classification. *IEEE Transactions on Medical Imaging*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; and Hassabis, D. 2015. Human-level control through deep reinforcement learning. *Nature*.

Oakden-Rayner, L.; and et al. 2020. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. *Proceedings of the National Academy of Sciences*, 117(48): 30678–30686.

Ozkan, A.; and Boix, G. 2020. Training across modalities for improved medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, 560–569.

Radford, A.; and et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. *ICML*.

Raghu, M.; and et al. 2021. Do Vision Transformers See Like Convolutional Neural Networks? *NeurIPS*.

Recht, B.; and et al. 2019. Do ImageNet Classifiers Generalize to ImageNet? *ICML*.

Sahoo, S.; and et al. 2022. A Neuro-Symbolic Framework for Interpretable Medical AI. *Artificial Intelligence in Medicine*, 130: 102343.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of ICCV*.

Shazeer, N.; and et al. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *ICLR*.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of ICLR*.

Singh, R.; and et al. 2024. Why Vision-Language Models Fail in Medical Domains. *Nature Machine Intelligence*.

Tschandl, P.; and et al. 2020. Human–Computer Collaboration for Skin Cancer Recognition. *Nature Medicine*.

Urooj, M.; Banerjee, A.; Shaikh, F.; Thakur, K.; and Gupta, S. 2025. MedXAI: A Retrieval-Augmented and Self-Verifying Framework for Knowledge-Guided Medical Image Analysis. *arXiv preprint arXiv:2512.10098*.

Volpi, M.; Zhang, Z.; and Chen, Y. 2018. Robustness of deep learning models in medical imaging: A survey. In *Proceedings of MICCAI*.

von Rueden, L.; and et al. 2021. Informed Machine Learning: Integrating Knowledge into Learning. *Frontiers in Artificial Intelligence*, 4: 57.

Wang, Y.; Xu, K.; and Lu, G. 2021. Enhancing medical image classification with domain-specific knowledge integration. In *Proceedings of ICML*.

Wu, Y.; Zhang, T.; and Holmes, J. 2022. Reinforcement learning for interpretable medical image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Xu, J.; and et al. 2023. Knowledge-Augmented Domain Generalization for Medical Imaging. *IEEE Transactions on Medical Imaging*.

Zhang, T.; and et al. 2024. Hybrid Expert Learning for Imbalanced Medical Datasets. *Medical Image Analysis*.

Zhang, Y.; and et al. 2023. BioMedCLIP: A Vision-Language Foundation Model for Biomedical Images and Text. *Nature Communications*.

Zhou, X.; and et al. 2023. Mixture-of-Experts for Multi-Organ Segmentation in Medical Imaging. *Medical Image Analysis*, 83: 102657.

Zhou, X.; Li, Y.; and Chen, P. 2022. Towards interpretable medical imaging with deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhou, Y.; and et al. 2022. Domain Generalization in Medical Imaging: A Survey. *Medical Image Analysis*, 83: 102667.

Zhou, Z.; and et al. 2021. A Survey of Ensemble Learning in Deep Neural Networks. *ACM Computing Surveys*.