

REN: Anatomically-Informed Mixture-of-Experts for Interstitial Lung Disease Diagnosis

Alec K. Peltekian, Halil Ertugrul Aktas, Gorkem Durak, Kevin Grudzinski,
Bradford C. Bemiss, Carrie Richardson, Jane E. Dematte, G. R. Scott Budinger,
Anthony J. Esposito, Alexander Misharin, Alok Choudhary, Ankit Agrawal, and Ulas Bagci

Abstract—Mixture-of-Experts (MoE) architectures achieve scalable learning by routing inputs to specialized subnetworks through conditional computation. However, conventional MoE designs assume homogeneous expert capability and domain-agnostic routing—assumptions that are fundamentally misaligned with medical imaging, where anatomical structure and regional disease heterogeneity govern pathological patterns. We introduce *Regional Expert Networks (REN)*, the first anatomically-informed MoE framework for medical image classification. REN encodes anatomical priors by training seven specialized experts, each dedicated to a distinct lung lobe or bilateral lung combination, enabling precise modeling of region-specific pathological variation. Multi-modal gating mechanisms dynamically integrate radiomics biomarkers with deep learning (DL) features extracted by convolutional (CNN), Transformer (ViT), and state-space (Mamba) architectures to weight expert contributions at inference. Applied to interstitial lung disease (ILD) classification on a 597-patient, 1,898-scan longitudinal cohort, REN achieves consistently superior performance: the radiomics-guided ensemble attains an average AUC of 0.8646 ± 0.0467 , a +12.5% improvement over the SwinUNETR single-model baseline (AUC 0.7685, $p = 0.031$). Lower-lobe experts reach AUCs

of 0.88-0.90, outperforming DL baselines (CNN: 0.76-0.79) and mirroring known patterns of basal ILD progression. Evaluated under rigorous patient-level cross-validation, REN demonstrates strong generalizability and clinical interpretability, establishing a scalable, anatomically-guided framework potentially extensible to other structured medical imaging tasks. Code is available on our GitHub <https://github.com/NUBagciLab/MoE-REN>.

Index Terms—Deep learning, interstitial lung disease, medical imaging, mixture-of-experts, radiomics

I. INTRODUCTION

In machine learning, *Mixture-of-experts (MoE)* architectures include multiple expert networks that focus on different aspects of complex data distributions. The success of MoE architectures stems from conditional computation; dynamically routing inputs to the most relevant experts while maintaining computational efficiency. However, existing MoE frameworks primarily operate under unconstrained optimization structures that assume homogeneous expert capabilities and domain-agnostic routing mechanisms. We argue that this approach is fundamentally misaligned with medical imaging. Anatomical structure, physiological constraints, and regional disease heterogeneity impose domain-specific specialization requirements that generic routing cannot capture effectively.

Unlike natural image classification, where semantic regions may be arbitrarily distributed, medical images exhibit well-defined anatomical structure where pathological patterns follow predictable regional distributions governed by underlying physiological processes [1], [2]. Traditional MoE systems cannot leverage these domain-specific priors, leading to suboptimal expert utilization and reduced interpretability, a critical requirement for clinical adoption [3].

Interstitial lung disease (ILD) diagnosis provides a concrete example of these challenges. ILD encompasses over 200 diverse pulmonary disorders. Accurate interpretation of high-resolution computed tomography (HRCT) scans is essential for distinguishing ILD subtypes and guiding clinical management [4], [5]. Since each anatomical region may exhibit varying disease manifestations, severity levels, and progression patterns within the same patient, conventional global analysis approaches struggle to provide both accurate diagnosis and interpretable outputs. Current DL methods for ILD classification treat the entire lung as a homogeneous unit, potentially diluting region-specific pathological signals and providing limited interpretability for clinical decision-making [6].

To this end, we introduce REN (Regional Expert Networks), a novel anatomically-informed MoE framework that addresses

A. K. Peltekian and A. Choudhary are with the Department of Computer Science, Northwestern University McCormick School of Engineering and Applied Science, Chicago, IL, United States. K. Grudzinski, B. C. Bemiss, J. E. Dematte, A. J. Esposito, G. R. S. Budinger, and A. Misharin are with the Division of Pulmonary and Critical Care Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, United States. C. Richardson is with the Division of Rheumatology, Northwestern University Feinberg School of Medicine, Chicago, IL, United States. G. R. S. Budinger and A. Misharin are also with the Simpson Querrey Lung Institute for Translational Science, Northwestern University Feinberg School of Medicine, Chicago, IL, United States. A. Choudhary and A. Agrawal are with the Department of Electrical and Computer Engineering, Northwestern University McCormick School of Engineering and Applied Science, Chicago, IL, United States. H. E. Aktas, G. Durak, and U. Bagci are with the Machine & Hybrid Intelligence Lab, Department of Radiology, Northwestern University Feinberg School of Medicine, Chicago, IL, United States.

This research was supported in part through a generous gift from K. Querrey and L. Simpson. We acknowledge the Quest high-performance computing facility and the Genomics Compute Cluster. G.R.S.B. was supported by a Chicago Biomedical Consortium grant, Northwestern University Dixon Translational Science Award, Simpson Querrey Lung Institute for Translational Science, the NIH (grant nos. P01AG049665, P01HL154998, U54AG079754, R01HL147575, R01HL158139, R01HL147290, R21AG075423 and U19AI135964), and the Veterans Administration (award no. I01CX001777). A.V.M. was supported by the NIH (grant nos. U19AI135964, P01AG049665, P01HL154998, U19AI181102, R01HL153312, R01HL158139, R01ES034350 and R21AG075423). A.C. was supported by NSF (grant no. OAC-2331329). A.A. was supported by the NIH (grant nos. U19AI135964 and R01HL158138), Simpson Querrey Lung Institute for Translational Science, and NSF (grant no. OAC-2331329). A.J.E. was supported by the NIH (grant no. L30HL149048). U.B. acknowledges the following grant: R01-HL171376. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. Corresponding author: U. Bagci (e-mail: ulas.bagci@northwestern.edu).

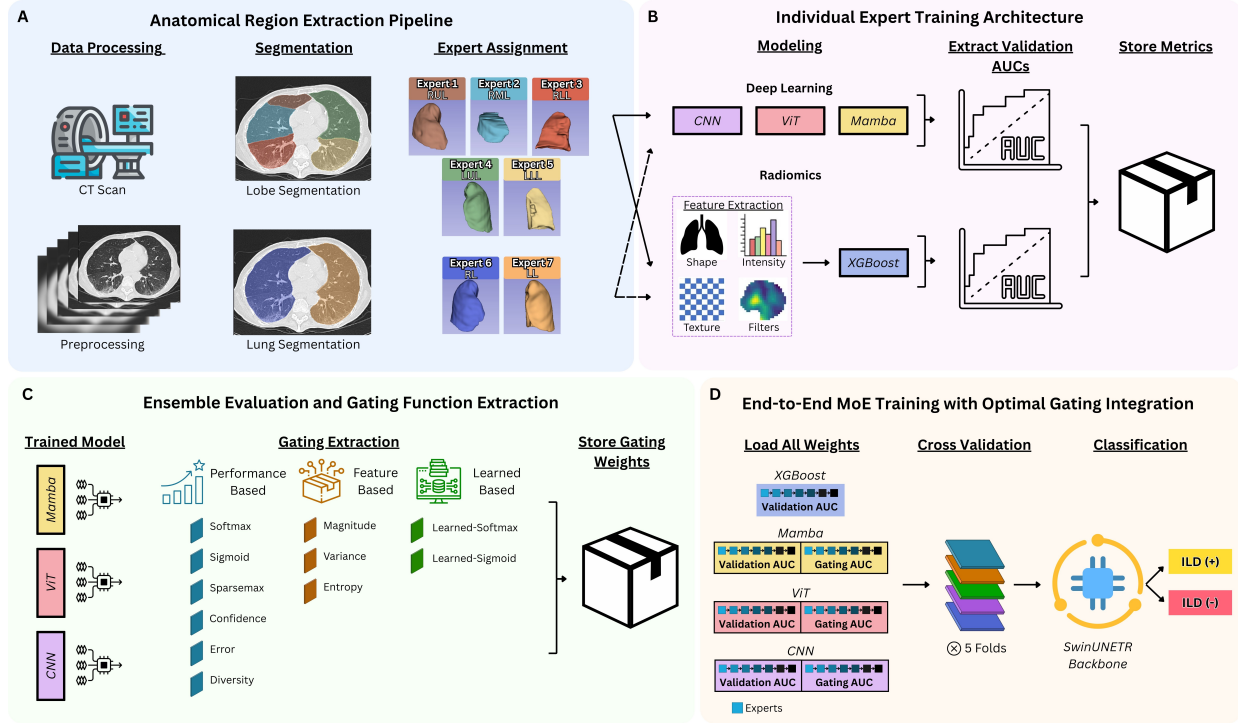


Fig. 1. Overview of the REN (Regional Expert Networks) framework. (A) Anatomical region extraction: preprocessing and lobe segmentation assign CT regions to seven experts (five lobes plus bilateral lungs). (B) Individual expert training: CNN, ViT, Mamba, and radiomics (XGBoost) experts are trained on masked inputs with validation AUCs recorded. (C) Gating function extraction: dynamic weighting strategies (performance-, feature-, and learned-based) are applied to expert outputs. (D) End-to-end MoE integration: expert weights and global SwinUNETR [7] features are fused for patient-level ILD classification.

the aforementioned limitations by embedding domain-specific constraints into expert specialization and routing mechanisms (Fig. 1). Unlike conventional MoE approaches where expert partitioning is learned implicitly, REN explicitly aligns expert specialization with anatomically defined lung regions, enabling interpretable regional modeling and structured expert routing tailored to ILD imaging patterns. Our novel approach demonstrates how anatomical priors can enhance MoE architectures. By training specialized experts for distinct anatomical regions and implementing multi-modal adaptive gating strategies, REN leverages inherent structural knowledge in medical applications while maintaining the scalability advantages of MoE systems. Through rigorous evaluation, we show that REN enables effective integration of both DL and traditional radiomics approaches. Particularly, radiomics-guided, newly designed gating strategies prove effective for capturing subtle pathological patterns in anatomically critical regions. **Our contributions** can be summarized as follows:

- **Novel Anatomically-Constrained MoE Architecture:** We introduce a new MoE framework specifically designed for medical imaging that incorporates anatomical structure as explicit constraints, addressing critical gaps in domain-specific expert routing. This is done by training seven lobe-specific experts that specialize in region-specific pathological pattern recognition through masked CT inputs.
- **Multi-Modal Gating Mechanisms:** We develop new gating strategies that integrate both DL features and traditional radiomics biomarkers, to dynamically balance

complementary information sources.

- **Comprehensive Architecture Evaluation:** We systematically compare three state-of-the-art architectures—a custom 3D CNN [8], a 3D Vision Transformer adapted from Dosovitskiy et al. [9], and (Vision) Mamba [10] architecture—across all anatomical regions within the lung using rigorous patient-level cross-validation, establishing architecture-specific regional specialization insights.
- **End-to-End Trainable Framework:** We implement a fully differentiable MoE system that enables joint optimization of expert networks and gating mechanisms while maintaining interpretability through expert contribution analysis.

II. RELATED WORK

A. Mixture of Experts in Deep Learning

MoE architectures leverage routing, expert diversity, and load balancing to enable scalable, efficient learning, as detailed in recent surveys [11]. While foundational for complex tasks, current MoE systems lack the domain-specific constraints critical for medical imaging, where anatomical structure guides disease patterns. A core challenge remains balancing expert specialization with computational cost. DeepSeek [12] advanced scalable MoE training through fine-grained expert partitioning and shared experts for generalization, whereas MoE++ [13] introduced zero-computation experts to enable efficient dynamic routing.

Extensions to vision and multi-modal settings have made MoE increasingly relevant to healthcare, while transformer-

based architectures have also been explored for ILD lesion analysis and weakly supervised segmentation tasks [14]. MoE-LLaVA [15] achieved sparse activation in vision-language models with only 3B active parameters, matching larger dense models via MoE-Tuning—a key step toward making MoE viable in resource-constrained clinical settings. Despite these advances, adapting MoE frameworks to incorporate anatomical priors and clinical semantics remains an open frontier for medical imaging.

Recent MoE research in medical imaging has also begun exploring domain-informed expert assignment and multi-modal integration. Wu et al. [3] proposed heterogeneous tissue experts for gigapixel whole-slide images and Jiang and Shen [16] introduced M4oE, a foundation model for multi-modal segmentation using MoE routing. These efforts highlight growing interest in adapting MoE to biomedical applications.

Our work differs by introducing an anatomically-constrained MoE tailored specifically for systemic sclerosis-associated ILD. Unlike prior approaches that define experts by tissue type, modality, or distributional shifts, we explicitly align experts with anatomical regions—mirroring radiological practice and enabling interpretable, region-level contributions. Furthermore, we integrate handcrafted radiomics features with deep expert outputs, yielding hybrid gating strategies that couple domain priors with multi-modal signals. This anatomically- and clinically-grounded orientation distinguishes our study from broader MoE developments and addresses interpretability demands unique to medical imaging.

B. Radiomics and Multi-Modal Integration

Radiomics remains a powerful methodology in medical imaging, particularly when integrated with DL techniques. By extracting quantitative features describing texture, shape, and intensity patterns, radiomics provides complementary information to the hierarchical representations learned by deep networks [17], [18]. This synergy continues to demonstrate significant effectiveness in improving diagnostic accuracy through multi-modal fusion. Traditional machine learning classifiers like XGBoost [19] maintain relevance for high-dimensional radiomics tasks due to their interpretability and capacity to model complex feature interactions. Nevertheless, current fusion strategies often overlook regional anatomical heterogeneity and differential diagnostic importance across tissue regions [20]. Anatomically-informed integration of radiomics with DL represents a promising yet underexplored frontier for enhancing both accuracy and clinical interpretability.

C. Anatomical Specialization and Regional Analysis

Integrating anatomical priors into DL models enhances performance and interpretability by emulating clinical radiology’s regional examination practices [21], [22]. Regional analysis proves particularly effective for pathologies with anatomically specialized patterns [23]. Conventional approaches that rely on whole-organ analysis often assume anatomical homogeneity and risk overlooking localized variation in pathology [24],

[25]. In pulmonary imaging, for example, global feature extraction across entire lung volumes can obscure localized disease signals, reducing both sensitivity and interpretability [26], [27]. This underscores the clinical need for anatomically informed models that capture region-specific patterns and align outputs with radiological workflows [28].

Recent advances incorporate anatomical knowledge through: (i) Denoising autoencoders and specialized architectures [29], (ii) Shape priors for segmentation accuracy [30], and (iii) Domain-informed constraints improving diagnostics [31]. Brain tumor segmentation exemplifies significant gains from anatomical priors [32]. This aligns naturally with MoE frameworks, where specialized subnetworks process distinct anatomical regions. The shift from generic to anatomically-informed expert routing represents a novel MoE application in medical imaging.

D. Deep Learning for ILD Diagnosis

Deep learning approaches for ILD analysis have grown substantially in recent years. Recent studies have explored CNN-based, multiple-instance-learning, radiomics-based, and transformer-based methods for tasks such as UIP prediction, lesion analysis, and disease pattern classification from high-resolution CT scans [6], [14]. Comprehensive reviews highlight the rapid development of artificial intelligence systems for ILD detection and disease pattern analysis from CT imaging [5]. Despite these advances, most existing approaches rely on global lung representations or task-specific pipelines rather than explicitly modeling anatomical regional specialization. REN instead introduces a mixture-of-experts framework that aligns expert specialization with anatomically defined lung regions.

E. Summary of Gaps and Motivation

Collectively, existing MoE frameworks lack the domain-specific constraints required for medical imaging; existing DL methods for ILD neglect regional anatomical structure; and existing radiomics–DL fusion strategies ignore regional heterogeneity. REN directly addresses all three limitations through an anatomically-constrained MoE design with multi-modal, region-aware gating.

III. METHODS

Our proposed anatomically-informed mixture-of-experts framework for ILD diagnosis (classification) consists of a four-stage pipeline that progressively builds from individual expert training to gating function extraction and finally to end-to-end MoE model creation. Fig. 1 represents REN in four stages: (A) anatomical region extraction through preprocessing and segmentation, (B) training of regional experts using CNN, ViT, Mamba, and radiomics models, (C) derivation of gating strategies to dynamically weight experts, and (D) end-to-end MoE integration with SwinUNETR for ILD classification. Global contextual features are extracted using SwinUNETR, a transformer-based medical imaging architecture that combines hierarchical Swin Transformer representations with UNet-style decoding for volumetric image analysis [7].

A. Dataset

Our study utilized a retrospective dataset from the Northwestern Scleroderma Registry comprising 597 patients with 1,898 longitudinal chest CT scans acquired between 2001 and 2023. This study was approved by the Northwestern University Institutional Review Board. All participants provided informed consent for inclusion in the Northwestern Scleroderma Registry at the time of enrollment. The cohort included 489 (81.9%) female patients with a mean age of 63.7 ± 12.7 years (range: 22.1-98.3 years). The study population was predominantly White (479 patients, 80.2%), with 84 (14.1%) Black, 20 (3.4%) Asian, and 14 (2.3%) other race patients. The dataset focused on systemic sclerosis-related conditions with the following disease subtype distribution: limited cutaneous systemic sclerosis (lcSSc): 284 (47.6%), diffuse cutaneous systemic sclerosis (dcSSc): 245 (41.0%), systemic sclerosis sine scleroderma (SSS): 28 (4.7%), with other subtypes (6.7%). Of the total cohort, 365 patients (61.1%) had confirmed ILD (via multidisciplinary consensus), forming the positive class for our binary classification task. Patients contributed an average of 3.2 scans each, reflecting longitudinal monitoring patterns typical in systemic sclerosis care.

B. Data Preprocessing and Evaluation Strategy

All CT scans, stored as NIFTI files, were processed using identical preprocessing pipelines, accompanied by corresponding lobe segmentation masks and binary classification labels. Lung and lobe segmentation were performed using the state-of-the-art *lungmask* algorithm [33], which provides robust automated segmentation of pulmonary lobes across diverse imaging protocols and pathological conditions. Our anatomical specialization approach focuses on seven distinct lung regions (automatically determined with *lungmask* segmentation): the five individual lobes (Left Upper Lobe - LUL, Left Lower Lobe - LLL, Right Upper Lobe - RUL, Right Middle Lobe - RML, Right Lower Lobe - RLL) plus two combined regions (Left Lung combining lobes 1 and 2, Right Lung combining lobes 3, 4, and 5) (Fig. 2). For evaluation, we implement strict patient-level cross-validation to prevent data leakage, where multiple scans per patient could artificially inflate performance if distributed across training and evaluation sets.

Each fold assigns 80% of patients to training and 20% to holdout, where the 20% holdout patients are unique to that fold. The holdout patients are then split 50/50 into validation (10% of total patients) and test (10% of total patients) sets (non-overlap). Across all 5 folds, every patient appears exactly once in the 20% holdout position, ensuring 50% of patients serve as validation data and 50% as test data across folds, with no patient contamination between sets.

Feature-space visualization: We qualitatively assessed regional separability by projecting radiomics feature vectors into two dimensions using t-SNE with standard settings and a fixed random seed. Embeddings were generated on the common set of scans across experts and are shown by anatomical region and by ILD status.

C. Regional Expert Networks (RENs) Architecture

Proposed REN modules specialize in distinct anatomical regions, each processing masked CT inputs corresponding to its assigned lung lobe or lung combination. This design ensures that learning is focused on region-specific pathological patterns rather than global averages. We implement several backbone variants: **CNN RENs** adopt a 3D convolutional architecture with three progressive blocks ($24 \rightarrow 48 \rightarrow 96$ filters, $3 \times 3 \times 3$ kernels), each followed by batch normalization, ReLU activation, and $2 \times 2 \times 2$ max pooling, before two fully connected layers ($256 \rightarrow 64$ units) with dropout 0.5. **ViT RENs** partition volumes into $8 \times 8 \times 8$ patches, embed them into 384 dimensions, and process through 6 transformer blocks (8 attention heads, MLP ratio 4.0, GELU activation) with learnable positional encodings. **Mamba RENs** similarly use $8 \times 8 \times 8$ patches (256-dimensional embeddings) processed through 4 Mamba blocks ($d_{\text{state}} = 64$, $d_{\text{conv}} = 4$, expansion factor 2) with SiLU activation. This multi-architecture design allows comparative evaluation of regional specialization across convolutional, attention-based, and state-space modeling structures. REN should therefore be viewed as a clinically grounded adaptation of mixture-of-experts modeling that incorporates anatomical priors and multi-modal feature guidance rather than as a new generic MoE learning algorithm.

D. Radiomics RENs

For the five individual lung lobes and two lung regions, we extract comprehensive radiomics feature sets using PyRadiomics [34]. The radiomics pipeline computes first-order statistics (mean, variance, skewness, kurtosis), shape-based features (volume, surface area, sphericity, compactness), and texture features including Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), Gray Level Dependence Matrix (GLDM), and Neighbourhood Grey Tone Difference Matrix (NGTDM) features. This feature set captures complementary aspects of lung tissue characterization. First-order features quantify density and heterogeneity, shape features capture anatomical alterations, and texture features detect spatial and microstructural patterns associated with ILD-related pathophysiology. We used a 25 Hounsfield Unit bin width for intensity discretization to balance sensitivity to tissue differences while maintaining robustness to noise across different scanner protocols. This results in 107 quantitative features per lung lobe that capture subtle patterns potentially missed by DL approaches while providing clinically interpretable measurements. XGBoost classifiers are trained for each lobe and lung using these extracted features.

E. Training Pipeline

Our end-to-end approach follows four sequential stages: (Stage 1) anatomical preprocessing and lobe segmentation; (Stage 2) regional expert training; (Stage 3) ensemble and gating extraction; and (Stage 4) end-to-end MoE optimization.

Stage 2 - Individual Regional Expert Network Training: In Stage 2, we trained specialized RENs for seven anatomical

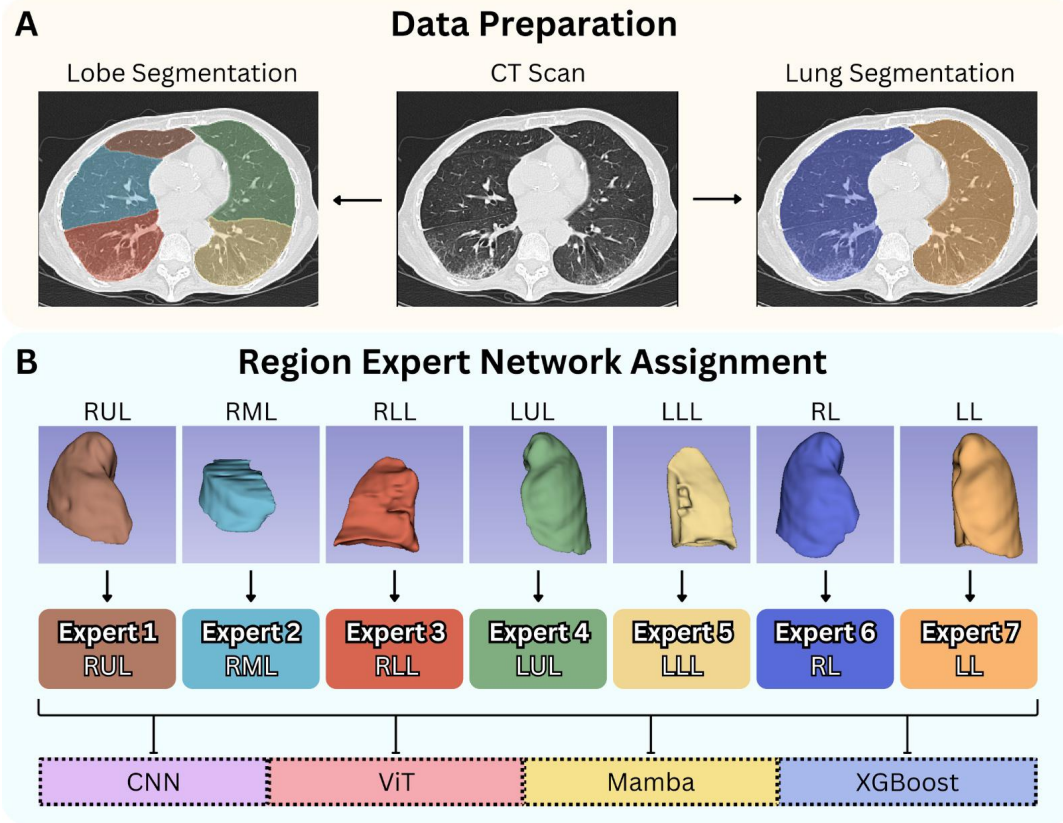


Fig. 2. Anatomical region extraction pipeline showing the process of generating masked inputs for each of the seven lung regions. The original CT scan is processed with lobe-specific segmentation masks to create region-focused inputs that enable expert specialization.

lung regions using three DL architectures (CNN, ViT, Mamba) and one radiomics model (XGBoost). Each REN received a masked CT input, isolating the target region via element-wise multiplication with its anatomical mask. Deep learning models were trained with architecture-specific hyperparameters and medical imaging-oriented augmentations (rotation, noise, affine transforms). CNN and Mamba used a learning rate of 1×10^{-4} , batch size of 4; ViT used a learning rate of 5×10^{-5} and a batch size of 2 due to memory demands. All models were optimized with AdamW and regularized with dropout, early stopping, and label smoothing. This stage yielded 28 RENs per fold (7 regions $\times$ 4 models), with validation AUCs used in Stage 3 ensemble integration.

Stage 3 – Staged Ensemble with Fixed Expert Weights.

In Stage 3, regional experts are combined using gating weights derived from validation data, including validation-AUC-based weighting rules and learned gating networks. Crucially, once selected, these gating weights are *fixed* and are not updated during subsequent inference or training. This staged design decouples expert learning from expert combination and preserves region-specific specialization.

Stage 3A - Basic Ensemble Evaluation: Stage 3 includes two sub-steps. Stage 3A establishes a simple ensemble baseline by combining the outputs of the regional experts rather than relying on any single model. The motivation for using an ensemble is that different experts capture complementary signals from distinct lung regions, so aggregating them can

reduce variance and improve overall stability. To formalize this, we compute a weighted prediction for each input, where the contribution of each expert is scaled by its validation AUC. This produces a single ensemble output that reflects both the predictions and the relative reliability of all experts. For a given input sample $\mathbf{x}$ with corresponding lobe mask $\mathbf{m}$, each expert E_k produces a prediction $\hat{y}_k(\mathbf{x})$, where $k \in \{1, 2, \dots, K\}$ represents the seven anatomical regions with $K = 7$. Equation (1) defines the weighted ensemble prediction:

$$\hat{y}_{\text{weighted}}(\mathbf{x}) = \sum_{k=1}^K w_k \cdot \hat{y}_k(\mathbf{x}), \quad (1)$$

where the normalized expert weights are computed as

$$w_k = \frac{\text{AUC}_k^{\text{val}}}{\sum_{j=1}^K \text{AUC}_j^{\text{val}}}. \quad (2)$$

Stage 3B - Candidate Gating Strategy Evaluation: In this stage, the proposed gating formulations are treated as candidate expert-weighting rules (heuristics) rather than as a single unified learning algorithm. Each strategy defines a different mechanism for assigning weights to regional experts, enabling us to study how alternative weighting principles influence the combination of region-specific predictions.

We considered three classes of candidate strategies: (1) performance-based weighting rules derived from validation

behavior, (2) feature-based weighting rules derived from intermediate representations, and (3) lightweight learned gating networks that produce input-dependent expert weights.

For each architecture and cross-validation fold, all candidate gating strategies were evaluated exclusively on the validation split. The strategy achieving the highest validation AUC was selected and then fixed for testing, ensuring no test-set leakage. Detailed mathematical formulations of the candidate weighting rules and learned gating architectures are provided in Appendix A.

Definition of Ensemble Variants: For clarity and reproducibility, we explicitly define all ensemble configurations used throughout this study:

- **Radiomics-Guided:** Expert weights are derived exclusively from validation AUCs of radiomics (XGBoost) regional models and applied to deep expert outputs.
- **Validation-Weighted:** Expert outputs are combined using fixed, global weights derived from normalized validation AUCs of the corresponding deep models.
- **CNN / ViT / Mamba Gated MoE:** End-to-end MoE models in which expert weights are input-dependent and learned via gating networks that are jointly optimized with the backbone.

All reported results correspond to Tables II–III, with computational and ablation analyses reported in Tables IV and V.

Stage 4 - End-to-End MoE Architecture Training: Stage 4 initializes expert and gating parameters using the optimal weights identified in Stage 3 and then jointly optimizes the backbone, experts, and gating in an end-to-end manner. Unlike Stage 3, which preserves independent regional training, Stage 4 employs a unified architecture that processes the full lung volume while adaptively weighting anatomically constrained expert contributions.

SwinUNETR was selected as the Stage 4 backbone due to its strong and well-documented performance in volumetric medical imaging, its ability to balance global contextual modeling with local spatial fidelity [7], and its widespread clinical adoption. The hierarchical Swin Transformer encoder captures long-range dependencies across lung regions, while the U-Net–style decoder preserves spatial structure, making it well suited for aggregating region-specific expert evidence into a unified representation. In our baseline experiments, SwinUNETR achieved higher full-lung AUC than alternative CNN-, ViT-, and Mamba-based backbones, motivating its use as the most competitive single-model reference. In addition, SwinUNETR benefits from large-scale medical imaging pre-training, which improves optimization stability and generalization in limited-data regimes. Importantly, REN is backbone-agnostic, and SwinUNETR is used here to provide a robust and reproducible reference for evaluating the proposed anatomically informed MoE formulation rather than to introduce architectural novelty.

Architecture Design: The proposed REN model processes the full CT volume $\mathbf{x} \in \mathbb{R}^{D \times H \times W}$ and lobe mask $\mathbf{m} \in \mathbb{R}^{D \times H \times W}$ using three main components. The SwinUNETR backbone extracts global features $\mathbf{F}_{\text{backbone}} = \text{SwinUNETR}(\mathbf{x})$, globally pooled to $\phi_{\text{global}} \in \mathbb{R}^{384}$. Seven expert extractors process masked inputs $\mathbf{x}_k = \mathbf{x} \odot \mathbf{m}_k$ to

produce features $\phi_k \in \mathbb{R}^{64}$, with attention weights α_k . Final expert weights are $g_k = w_k \alpha_k$, normalized such that $\sum_{k=1}^K g_k = 1$. Weighted expert features $\phi_{\text{expert}} = \sum_{k=1}^K g_k \phi_k$ are concatenated with global features to form $\phi_{\text{combined}} = [\phi_{\text{global}}; \phi_{\text{expert}}] \in \mathbb{R}^{448}$, classified by an MLP as $y_{\text{pred}} = \text{MLP}(\phi_{\text{combined}})$.

Weight Initialization Strategies: Stage 4 initializes and compares seven strategies: four based on validation AUC weighting from Stage 2 (radiomics, CNN, ViT, Mamba) and three based on the optimal learned gating from Stage 3 ($\mathbf{w}_{\text{CNN}}^*$, $\mathbf{w}_{\text{ViT}}^*$, $\mathbf{w}_{\text{Mamba}}^*$). For each fold, the selected initialization weights were normalized within the anatomical hierarchy before end-to-end optimization. Full formulations are provided in Appendix A.

Training Configuration: The end-to-end SwinUNETR MoE is trained with AdamW using differentiated learning rates (10^{-5} for backbone, 10^{-4} for MoE) and cosine annealing scheduling. The total loss is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{gating}} \mathcal{L}_{\text{gating}} + \lambda_{\text{weight}} \mathcal{L}_{\text{weight}} + \lambda_{\text{diversity}} \mathcal{L}_{\text{diversity}}, \quad (3)$$

with hyperparameters $\lambda_{\text{gating}} = 0.005$, $\lambda_{\text{weight}} = 0.005$, and $\lambda_{\text{diversity}} = 0.01$ selected via validation.

F. Scope of Anatomical Priors

We operationalize anatomical knowledge as an inductive bias through region-constrained expert assignment, rather than as a formal probabilistic prior. Specifically, expert specialization is enforced through anatomically defined region decomposition and masked inputs, reflecting radiological practice.

G. Summary of Methodological Innovations

First, we introduce RENs, the first anatomically informed MoE architecture in medical imaging, specifically in lung diseases. Expert assignment is guided explicitly by lobe- and lung-level segmentation rather than unconstrained routing. This ensures that expert specialization mirrors radiological practice and directly encodes anatomical priors into the model.

Second, we propose a **novel multi-modal gating framework** that integrates three complementary strategies: (i) performance-based weighting derived from validation AUCs, (ii) feature-based weighting using intermediate representations, and (iii) lightweight learned gating networks. This design enables dynamic, input-adaptive expert selection while avoiding the interpretability and load-balancing limitations of conventional MoE routing.

Third, we incorporate **radiomics-guided expert models**, providing handcrafted, pathology-aware biomarkers that complement deep learning features. By embedding radiomics in both regional experts and gating mechanisms, our framework enhances interpretability and sensitivity to subtle, anatomically localized disease patterns.

Role of Radiomics in REN: We acknowledge that radiomics–deep learning fusion has been explored in prior ILD and thoracic imaging studies [4]. In our current study, our contribution lies in how radiomics is used within REN: radiomics features are not fused at the representation level, but serve as

region-specific reliability signals that guide expert weighting within an anatomically constrained MoE framework. This design preserves modularity between handcrafted and learned representations while enabling region-aware interpretability through anatomically grounded expert weighting.

Collectively, these innovations constitute a domain-specific adaptation of MoE tailored for ILD classification. REN is not a simple application of existing MoE systems; it is a redesign that leverages anatomical priors, radiomics integration, and hybrid gating to produce clinically interpretable predictions.

IV. EXPERIMENTS AND RESULTS

We evaluated REN through comprehensive patient-level cross-validation experiments across three DL architectures and multiple gating strategies, using two anatomical configurations (five individual lobes; five lobes plus two bilateral regions). Table I maps each claimed contribution to its corresponding experimental evidence.

Model-Specific Hyperparameters: All DL experts were trained using model-specific hyperparameters optimized for medical imaging tasks. CNN and Mamba experts used learning rate 1×10^{-4} , batch size 4, while ViT experts used learning rate 5×10^{-5} , batch size 2. All models used AdamW optimizer with weight decay 1×10^{-5} . Early stopping was applied with a patience of 15 epochs. Learning rate scheduling used ReduceLROnPlateau for CNNs and CosineAnnealingLR for Mamba and ViT models. Dropout rates were 0.5 for CNNs, 0.2 for Mamba and ViT. Label smoothing was 0.05 for CNN, 0.2 for ViT and Mamba.

Computational Environment and Reproducibility: All experiments were conducted on NVIDIA A100-SXM4-80GB GPUs, using CUDA 12.1 and driver version 525.85.12. The software environment consisted of Python 3.8.20, PyTorch 2.4.1+cu121, MONAI 1.3.2, NumPy 1.24.4, scikit-learn 1.3.2, Pandas 2.0.3, XGBoost 2.1.4, and pyradiomics v3.1.0. Mamba models used mamba-ssm 2.2.2 for state-space computations.

For reproducibility, all random number generators were initialized with seed=42 using a seeding function that sets Python’s random module, NumPy, PyTorch CPU and GPU generators, and enables CUDA deterministic operations. Cross-validation splits used fold-specific seeds to ensure consistent patient partitioning across all experiments.

A. Stage 2: Individual Expert Performance

Architecture Performance Comparison: Radiomics-based MoE models outperformed neural networks based MoEs in most region-fold comparisons (5 lobes $\times$ 5 folds = 25): 23/25 vs. Mamba, 24/25 vs. ViT, and 17/25 vs. CNN, with mean AUC differences of 0.117 ± 0.044 (95% CI: [0.073, 0.162]), 0.193 ± 0.053 (95% CI: [0.140, 0.246]), and 0.069 ± 0.046 (95% CI: [0.022, 0.115]), respectively (Fig. 3). This performance advantage likely stems from radiomics’ domain-specific features (n=107), which encode texture, shape, and intensity, offering compact, pathology-aware representations and reducing overfitting risks on limited data. Among neural networks, CNNs showed the closest performance to radiomics, suggesting convolutional filters effectively model local lung structures.

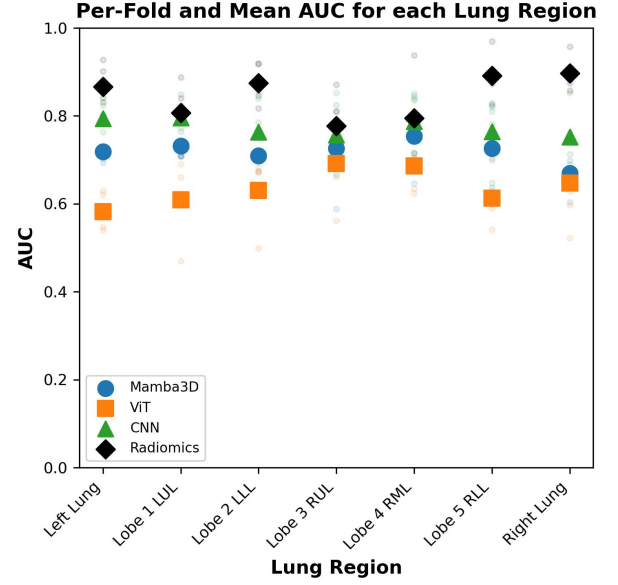


Fig. 3. Mean AUC per lung region across folds for each architecture. Radiomics experts achieved the highest regional performance, particularly in the lower lobes, while CNN followed closely behind. Mamba and ViT showed more variability across regions.

More advanced models, ViT and Mamba, originally optimized for non-medical domains, lagged behind—highlighting the need for adaptation when applying transformer-style models to clinical imaging [14]. Although the CNN expert contains substantially more parameters (42.6M) than the ViT (0.6M) and Mamba (0.7M) experts, all models were trained under identical data splits, preprocessing, and cross-validation conditions. Therefore, the observed performance differences reflect architectural inductive biases rather than differences in training protocol.

Regional Specialization Patterns: Model performance mirrored known ILD progression patterns, with highest AUCs in the left and right lower lobes (LLL/RLL), consistently exceeding 0.85 in the radiomics-based MoE approach. This aligns with clinical evidence that ILD typically begins in the lung bases due to gravitational and mechanical stress. Radiomics was particularly sensitive to early fibrotic features—reticulation and ground-glass opacities—in these regions. In contrast, reduced performance in the upper lobes and right middle lobe (RML) across all methods reflects the later onset of ILD in these areas. These results suggest the models capture the true anatomical progression of ILD, rather than spurious spatial patterns.

B. Stage 3: Learned and Gating Strategy Evaluation

Learned vs. Static Gating Performance: Learned gating consistently outperforms static weighting across architectures within the gating strategy evaluation, improving model stability through adaptive expert selection. This validates dynamic gating as essential for medical MoE architectures, where input-dependent routing captures patient-specific anatomical variations while maintaining predictable behavior across diverse cases.

TABLE I
MAPPING OF REN CONTRIBUTIONS TO EXPERIMENTAL EVALUATIONS.

Contribution	Experimental Variant	Evidence
Anatomical expert specialization	Regional CNN, ViT, and Mamba experts trained per lung region	Tables II, III
Radiomics-guided expert weighting	Radiomics-guided ensemble weighting	Tables II, III
Staged ensemble vs end-to-end MoE	Validation-weighted ensemble vs gated MoE	Tables II, III
Expert routing behavior	Routing diversity and usage metrics under regularization	Table V
Computational feasibility	Training time, memory, and parameter analysis	Table IV

TABLE II
FINAL PERFORMANCE USING FIVE INDIVIDUAL LUNG LOBES: BASELINE VS. VALIDATION-WEIGHTED VS. END-TO-END GATED MOE. PERCENTAGE CHANGE IS RELATIVE TO THE SWINUNETR BASELINE (AUC 0.7685).

Method	AUC $\pm$ SD [95% CI]	Change	p-value
Baseline			
SwinUNETR (Baseline)	0.7685 $\pm$ 0.0759 [0.674, 0.863]	–	–
CNN (Baseline)	0.7584 $\pm$ 0.0919 [0.667, 0.850]	–	–
Mamba (Baseline)	0.6775 $\pm$ 0.0454 [0.632, 0.723]	–	–
ViT (Baseline)	0.6535 $\pm$ 0.0356 [0.618, 0.689]	–	–
Validation-Weighted			
CNN Weighted	0.8033 $\pm$ 0.0472 [0.745, 0.862]	+4.5%	0.176
ViT Weighted	0.8015 $\pm$ 0.0453 [0.745, 0.858]	+4.3%	0.179
Mamba Weighted	0.7995 $\pm$ 0.0466 [0.742, 0.857]	+4.0%	0.213
Radiomics-Guided	0.8646 $\pm$ 0.0467 [0.806, 0.923]	+12.5%	0.031
End-to-End Gated MoE			
CNN Gated MoE	0.7760 $\pm$ 0.0689 [0.691, 0.862]	+1.0%	0.819
ViT Gated MoE	0.8010 $\pm$ 0.0430 [0.748, 0.854]	+4.2%	0.166
Mamba Gated MoE	0.8164 $\pm$ 0.0393 [0.768, 0.865]	+6.2%	0.053

TABLE III
FINAL PERFORMANCE COMPARISON USING 5 LOBES + 2 BILATERAL LUNG REGIONS (7 TOTAL): BASELINE VS. WEIGHTED MOE VS. GATED MOE. PERCENTAGE CHANGE IS RELATIVE TO THE SWINUNETR BASELINE (AUC 0.7685).

Method	AUC $\pm$ SD [95% CI]	Change	p-value
Baseline			
SwinUNETR (Baseline)	0.7685 $\pm$ 0.0759 [0.674, 0.863]	–	–
CNN (Baseline)	0.7584 $\pm$ 0.0919 [0.667, 0.850]	–	–
Mamba (Baseline)	0.6775 $\pm$ 0.0454 [0.632, 0.723]	–	–
ViT (Baseline)	0.6535 $\pm$ 0.0356 [0.618, 0.689]	–	–
Validation-Weighted			
CNN Weighted	0.7960 $\pm$ 0.0512 [0.732, 0.860]	+3.6%	0.382
ViT Weighted	0.8021 $\pm$ 0.0577 [0.730, 0.874]	+4.4%	0.297
Mamba Weighted	0.7983 $\pm$ 0.0490 [0.737, 0.860]	+3.9%	0.351
Radiomics-Guided	0.8523 $\pm$ 0.0430 [0.796, 0.908]	+10.9%	0.014
End-to-End Gated MoE			
CNN Gated MoE	0.7778 $\pm$ 0.0545 [0.710, 0.846]	+1.2%	0.442
ViT Gated MoE	0.7946 $\pm$ 0.0328 [0.754, 0.835]	+3.4%	0.242
Mamba Gated MoE	0.7950 $\pm$ 0.0486 [0.735, 0.855]	+3.4%	0.278

C. Stage 4: End-to-End Integration Results

Tables II and III compare final test performance between weighted ensemble approaches, end-to-end MoE architectures, and baseline methods using different anatomical configurations.

Recommended configuration: The highest diagnostic performance was achieved by the *staged radiomics-guided REN ensemble using the five-lobe configuration* (average AUC 0.8646), which we therefore consider the primary accuracy-oriented configuration. The end-to-end gated MoE variants are included primarily to analyze expert routing and specialization under joint optimization rather than as the recommended deployment configuration. In traditional computer vision, joint optimization allows experts to dynamically partition the feature space. However, our ablation studies reveal that in the medical imaging problem we address, joint optimization causes partially overlapping expert representation that actively degrades the strict anatomical specialization (e.g., lobe-specific pathological patterns) required for ILD diagnosis. Hence, while end-to-end MoE is theoretically appealing, preserving independently trained experts via staged ensembling is fundamentally better suited for anatomically constrained medical tasks.

Anatomical Configuration Comparison: We evaluated two anatomical expert configurations: (1) five individual lung lobes only, and (2) seven regions including five individual lobes plus bilateral lung combinations. The five-lobe configuration achieved superior performance, with the radiomics-guided ensemble reaching an average AUC of 0.8646 compared to 0.8523 for the seven-region configuration. This 1.4%

TABLE IV
TRAINING AND INFERENCE COMPLEXITY ACROSS MODELS

Model	Training Time (GPU-hours)	Peak Memory (GB)	Parameters (Millions)
CNN Expert (single region)	0.82	1.94	42.6
ViT Expert (single region)	0.82	1.66	0.6
Mamba Expert (single region)	0.81	0.27	0.7
SwinUNETR Baseline	0.91	12.98	62.2
MoE Pipeline (CNN expert)	5.10	19.39	63.6
MoE Pipeline (ViT expert)	4.98	19.39	63.6
MoE Pipeline (Mamba expert)	5.04	19.39	63.6
MoE Pipeline (Radiomics expert)	5.07	19.39	63.6
Gated MoE (end-to-end)	5.02	19.74	64.7
Inference Performance			
Model	Inference Time (ms)	Samples/sec	
CNN / ViT / Mamba Experts	46–48	21+	
Radiomics-Guided REN Ensemble	62	16	

improvement suggests that including bilateral lung experts may introduce redundancy that dilutes the specialized knowledge of individual lobe experts.

Primary Performance and Computational Burden: The radiomics-guided ensemble achieved the highest performance using the five-lobe configuration (average AUC 0.8646). This represents a statistically significant 12.5% improvement over baseline SwinUNETR ($p = 0.031$, paired t-test across CV folds). This demonstrates the complementary value of engineered features when combined with anatomically informed expert weighting. All MoE variants meet typical clinical deployment requirements. Inference times are 46–48 ms per patient (21+ samples/s) for DL approaches and 62 ms for radiomics-guided ensembles. Model sizes (237–247 MB;

TABLE V
ABLATION STUDY OF GATING REGULARIZATION EFFECTS ON EXPERT
BEHAVIOR OF ALL EXPERTS

Configuration	Mean Active Experts	Usage Variance	Mean Entropy	Cosine Similarity
Baseline (Full Loss)	3.10	237.2	1.43	0.49
No Regularization	2.25	920.2	0.85	0.83
Remove Entropy	2.58	620.2	1.00	0.41
Remove Load Balancing	3.01	290.3	1.35	0.53
Remove Diversity	3.34	183.2	1.49	0.82
Entropy Weight = 0.001	2.80	532.8	1.12	0.44
Entropy Weight = 0.002	2.69	673.4	1.07	0.52
Entropy Weight = 0.01	3.76	97.5	1.68	0.54
Entropy Weight = 0.02	5.12	60.7	1.79	0.60
Load Balance Weight = 0.001	3.36	238.0	1.44	0.52
Load Balance Weight = 0.005	3.49	142.2	1.58	0.48
Load Balance Weight = 0.02	3.48	226.1	1.45	0.57
Load Balance Weight = 0.05	3.11	305.4	1.38	0.61
Diversity Weight = 0.001	3.03	223.6	1.38	0.83
Diversity Weight = 0.005	3.05	256.7	1.40	0.60
Diversity Weight = 0.02	2.61	572.2	1.21	0.38
Diversity Weight = 0.05	2.52	721.3	1.12	0.29

62–65 M parameters) remain manageable for standard clinical hardware. These speeds far exceed human interpretation time, enabling real-time diagnostic support. Importantly, weighted ensembles consistently outperform end-to-end gated counterparts. This validates our training strategy: staged ensembling preserves specialized anatomical knowledge that joint optimization obscures. The modest computational overhead of radiomics integration (≈ 16 ms) is negligible relative to its 12.5% performance gain. This establishes the hybrid approach as both computationally feasible and diagnostically superior for clinical deployment.

Detailed training/inference time, memory usage, and parameter counts for all expert and MoE variants are reported in Table IV. For clarity, the reported complexity for CNN, ViT, and Mamba models corresponds to *per-expert training*, while the MoE pipeline rows represent the computational cost of the *full REN ensemble*, including all regional experts and gating components. Although staged MoE pipelines incur higher total training cost due to independent expert training, peak memory usage remains below 20 GB and per-model training time is comparable to the SwinUNETR baseline, supporting practical clinical deployment.

Clinical Translation and Significance: The 12.5% improvement could translate to better identification of early-stage ILD cases, where timely diagnosis is critical for patient outcomes. The statistical significance of only the radiomics-guided approach suggests that while other methods show numerical improvements, they may not provide reliable enough enhancement for clinical implementation.

Ablation of Gating Regularization Terms: Table V analyzes the effect of individual gating regularizers on expert utilization and redundancy. Removing all regularization leads to severe expert collapse, characterized by high usage variance and strong feature similarity between experts. Entropy regularization controls the number of active experts and prevents overconfident routing, while load-balancing regularization reduces expert usage variance and promotes equitable expert

participation. Diversity regularization plays a critical role in reducing redundancy, as removing it substantially increases cosine similarity between expert representations. Together, these components stabilize expert behavior, balancing specialization, coverage, and interpretability.

Radiomics Analysis: Comprehensive ablation studies reveal texture features (GLCM, GLRLM) drive performance in lower lobes where ILD initiates (Fig. 4). The embeddings corroborate the quantitative results: texture families (GLCM/GLRLM) yield tight, separable lower-lobe clusters, while upper-lobe geometry remains fragmented, explaining regional AUC differences. Lower lobe clustering patterns align with clinical knowledge of disease progression, while the approach’s consistency across cross-validation folds ensures reliability across diverse clinical scenarios—critical for real-world deployment. The 7.6-8.1% improvement over individual ensembles demonstrates tangible workflow benefits. This hybrid approach combines radiomics interpretability with DL’s pattern recognition, providing both diagnostic accuracy and explainable insights. Regional performance breakdowns enable targeted radiologist attention and support differential diagnosis based on anatomical distribution patterns, aligning with natural radiological evaluation practices and facilitating clinical integration.

V. DISCUSSION AND CONCLUDING REMARKS

REN demonstrates a principled and effective approach to anatomically-informed MoE architectures in medical imaging. By aligning expert specialization with lobe-level anatomy and integrating radiomics biomarkers via multi-modal gating, REN consistently outperforms both single-model baselines and unconstrained ensemble approaches for ILD classification. The 12.5% AUC gain over the SwinUNETR baseline is statistically significant and clinically meaningful, as improved sensitivity at this margin could translate directly to earlier identification of ILD—where timely diagnosis is critical for slowing disease progression.

Three key insights emerge from the experimental results. First, cross-architecture comparison reveals region-specific performance profiles: CNN and radiomics experts excel in lower lobes where ILD initiates, while transformer-based architectures offer complementary sensitivity in regions with more diffuse involvement. This suggests that heterogeneous expert assignment—selecting the architecture best suited to each anatomical region—may yield further gains over uniform architectural deployment. Second, staged ensembling consistently and substantially outperforms end-to-end joint optimization. When experts are jointly trained, shared gradients erode the anatomical specialization that makes each expert informative; the ablation data in Table V make this collapse explicit. Preserving independently trained regional experts and combining them via fixed gating is the correct design choice for anatomically-constrained tasks. Third, the performance advantage of radiomics-guided weighting arises not from radiomics features alone, but from their synergy with lobe-specific expert assignment. Whole-lung radiomics models perform substantially worse than lobe-specific ra-

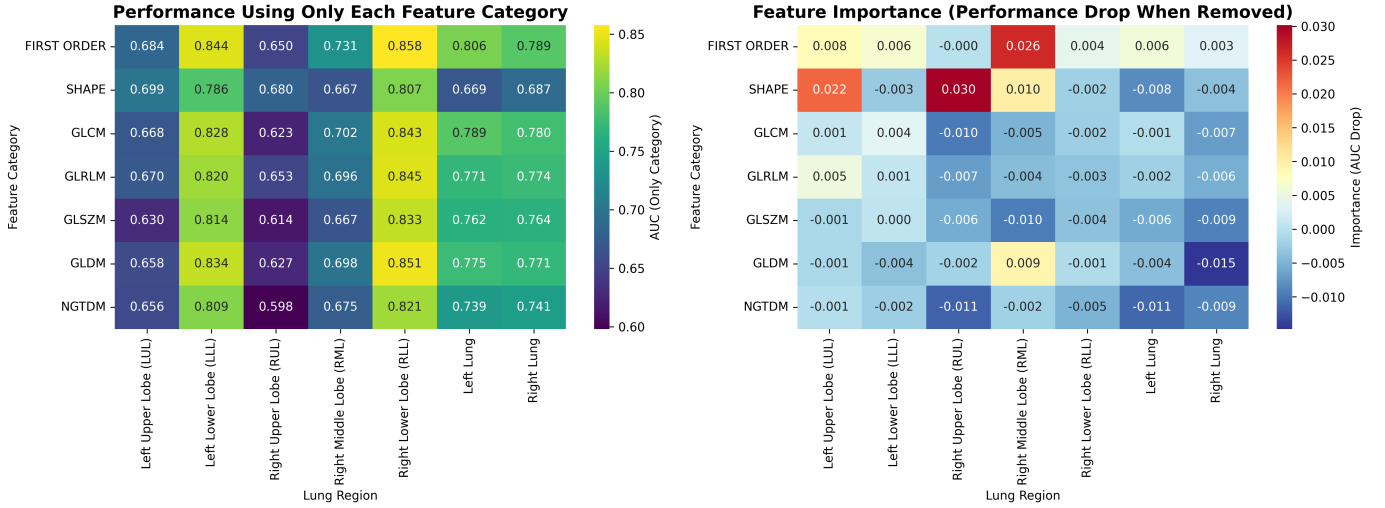


Fig. 4. Radiomics feature category ablation analysis showing (left) performance using individual feature categories across all lung regions and (right) feature importance measured by AUC drop when each category is removed. Texture features demonstrate particular strength in lower lobe regions, with GLCM achieving 0.828 AUC in Left Lower Lobe and GLRLM reaching 0.845 in Right Lower Lobe.

diomics experts, confirming that anatomical decomposition is a prerequisite for effective handcrafted feature utilization.

Anatomical grounding also confers direct interpretability benefits. Gating weights provide per-patient, per-region evidence maps that enable radiologists to identify which lung lobes drove the model’s prediction—a form of spatial attribution that aligns naturally with existing reporting workflows. Unlike saliency maps or attention visualizations that require post-hoc analysis, REN’s lobe-level contributions are intrinsic to the model architecture and operationally interpretable without additional computation.

Dense expert activation is a deliberate design choice that distinguishes REN from sparsely-activated MoE architectures. Clinical radiology requires systematic assessment of all anatomically relevant regions; enforcing hard sparsity would suppress informationally relevant experts for computational savings that are unnecessary given REN’s already-feasible 62 ms inference latency. Soft sparsity through gating down-weighting uninformative regions per patient achieves an appropriate balance between anatomical completeness and computational efficiency.

For clinical deployment, REN is intended as a decision-support module rather than an autonomous diagnostic system. It processes routine chest CT scans, performs automated lobe segmentation and regional analysis, and outputs a global ILD confidence label with calibrated per-lobe expert contributions. Practical deployment requires robust handling of protocol variability and common confounders (infection, edema, motion, emphysema), quality-control checks, and a human-in-the-loop workflow. Before clinical use, multi-center external validation and prospective evaluation are necessary, along with governance aligned with privacy and regulatory standards.

Model compression for deployment is tractable due to REN’s staged design. Knowledge distillation—training a compact student network using REN ensemble soft predictions and region-level intermediate features—is the most natural compression strategy. Expert pruning (removing experts with

consistently low gating weights), top-k regional routing, and post-training quantization offer further options for latency and memory reduction without sacrificing diagnostic performance.

Limitations: Several limitations constrain the present findings. The dataset originates from a single institution and focuses exclusively on systemic sclerosis-related ILD, which may limit generalizability to other ILD subtypes and patient populations. The 2001–2023 temporal range introduces variability in imaging protocols and scanner hardware; while this serves as a natural domain-shift stress test (1,898 scans across 11 affiliated hospitals over 22 years), it is not a substitute for prospective multi-center validation. Reliance on automated lungmask segmentation may propagate boundary errors throughout the pipeline; future work will evaluate sensitivity to segmentation quality through controlled perturbation experiments. The current binary ILD classification does not capture heterogeneity across subtypes, and fixed anatomical expert assignment may limit adaptability to disease-specific spatial patterns (e.g., mid-upper lung involvement in hypersensitivity pneumonitis). The ensemble’s total computational cost—from training and deploying seven independent regional networks—may also present scalability challenges in resource-constrained settings.

Although cross-institutional data sharing currently precludes additional external validation, ongoing multi-institutional collaborations are planned to evaluate REN across heterogeneous scanners, patient demographics, and ILD etiologies under domain shift.

Future Directions: We identify four priority directions for future work: (1) multi-institutional validation across demographics, scanner types, and ILD etiologies; (2) expansion to multi-class ILD subtype classification; (3) knowledge distillation strategies that compress the ensemble while preserving the observed 12.5% performance gain; and (4) adaptive expert assignment driven by pathological distribution rather than fixed anatomical regions. Incorporating graph-based MoE formula-

tions [35] could enable relational routing that captures anatomical adjacency and disease-spread dynamics across lobes. Extension to vision-language domains—integrating imaging with radiology reports via multimodal MoE architectures such as MedMoE [36], represents a natural next step toward clinically actionable AI that reasons jointly over images and text.

APPENDIX A METHODS AND MATHEMATICAL FORMULATIONS

This appendix provides detailed mathematical formulations and architectural specifications for the gating strategies and expert-weighting mechanisms used in REN. These formulations are included to ensure full reproducibility while keeping the main manuscript focused on the core architectural design and experimental evaluation.

A. Candidate Gating Strategy Formulations

1) *Enhanced AUC Normalization*: Various normalization schemes, such as softmax, sigmoid, and sparsemax are applied to validation AUC scores in (4)–(6),

$$g_k^{\text{val-auc-softmax}} = \frac{\exp(\text{AUC}_k^{\text{val}})}{\sum_{j=1}^K \exp(\text{AUC}_j^{\text{val}})}, \quad (4)$$

$$g_k^{\text{val-auc-sigmoid}} = \frac{\sigma(\text{AUC}_k^{\text{val}})}{\sum_{j=1}^K \sigma(\text{AUC}_j^{\text{val}})}, \quad (5)$$

$$g_k^{\text{val-auc-sparsemax}} = \max(\text{AUC}_k^{\text{val}} - \tau, 0), \quad (6)$$

where $\sigma(x) = 1/(1+e^{-x})$ and τ ensures unity sum for sparsemax. Softmax emphasizes small differences between AUC scores by producing dense probability distributions, sigmoid compresses the range of values to create smoother weightings, and sparsemax projects onto the simplex with exact zeros to encourage sparsity. These approaches differ in how strongly they amplify, smooth, or prune expert contributions.

In sparsemax normalization, the threshold τ is computed as the unique value that ensures $\sum_k \max(\text{AUC}_k^{\text{val}} - \tau, 0) = 1$, following the standard sparsemax formulation.

2) *Dynamic Performance Gating*: Sample-dependent weights based on prediction characteristics are defined in (7)–(9),

$$g_k^{\text{confidence}}(\mathbf{x}) = \frac{\exp(2|\hat{y}_k(\mathbf{x}) - 0.5|)}{\sum_{j=1}^K \exp(2|\hat{y}_j(\mathbf{x}) - 0.5|)}, \quad (7)$$

$$g_k^{\text{error}}(\mathbf{x}) = \frac{\exp(-|\hat{y}_k(\mathbf{x}) - y|)}{\sum_{j=1}^K \exp(-|\hat{y}_j(\mathbf{x}) - y|)}, \quad (8)$$

$$g_k^{\text{diversity}} = \frac{\exp(1 - \bar{\rho}_k)}{\sum_{j=1}^K \exp(1 - \bar{\rho}_j)}, \quad (9)$$

where $\bar{\rho}_k$ represents the average correlation between expert k and all others. Confidence, error, and diversity each address complementary aspects of expert reliability. Confidence weighting gives more influence to experts that make decisive predictions far from uncertainty, error weighting prioritizes experts that consistently align with the ground truth, and diversity weighting favors experts that provide outputs less correlated with the others.

3) *Statistical Feature Analysis*: Feature-vector-based weighting strategies are shown in (10)–(12),

$$g_k^{\text{magnitude}}(\mathbf{x}) = \frac{\exp(\|\phi_k(\mathbf{x})\|_2)}{\sum_{j=1}^K \exp(\|\phi_j(\mathbf{x})\|_2)}, \quad (10)$$

$$g_k^{\text{variance}}(\mathbf{x}) = \frac{\exp(\text{Var}(\phi_k(\mathbf{x})))}{\sum_{j=1}^K \exp(\text{Var}(\phi_j(\mathbf{x})))}, \quad (11)$$

$$g_k^{\text{entropy}}(\mathbf{x}) = \frac{\exp(H_k(\mathbf{x}))}{\sum_{j=1}^K \exp(H_j(\mathbf{x}))}, \quad (12)$$

where $\phi_k(\mathbf{x})$ represents extracted features and $H_k(\mathbf{x}) = -\sum_{b=1}^{20} p_{k,b} \log(p_{k,b} + \epsilon)$ is entropy from 20-bin histograms. Magnitude weighting emphasizes experts with stronger overall feature activations, variance weighting favors experts whose features display greater spread, and entropy weighting highlights experts with more uniform distributions. These methods differ in whether they capture signal strength, variability, or uncertainty, providing complementary perspectives for weighting experts.

B. Learned Gating Network Architectures

1) *Architecture-Specific Gating Formulations*: The CNN gating network applies multi-layer perceptron processing of FC1 features as in (13)–(15),

$$\mathbf{h}_1 = \text{ReLU}(\mathbf{W}_1 \bar{\phi}^{\text{FC1}}(\mathbf{x}) + \mathbf{b}_1), \quad 64 \text{ units}, \quad (13)$$

$$\mathbf{h}_2 = \text{ReLU}(\mathbf{W}_2 \mathbf{h}_1 + \mathbf{b}_2), \quad 32 \text{ units}, \quad (14)$$

$$\mathbf{z}(\mathbf{x}) = \mathbf{W}_3 \mathbf{h}_2 + \mathbf{b}_3, \quad 7 \text{ outputs}. \quad (15)$$

The Mamba gating network incorporates multi-head attention in (16)–(18),

$$\mathbf{h}_{\text{proj}} = \mathbf{W}_{\text{proj}} \bar{\phi}^{\text{pre-cl}}(\mathbf{x}) + \mathbf{b}_{\text{proj}}, \quad (16)$$

$$\mathbf{h}_{\text{attn}} = \text{MultiHeadAttention}(\mathbf{h}_{\text{proj}}, \mathbf{h}_{\text{proj}}, \mathbf{h}_{\text{proj}}), \quad (17)$$

$$\mathbf{z}(\mathbf{x}) = \text{MLP}(\mathbf{h}_{\text{attn}}). \quad (18)$$

The ViT gating network applies transformer-style processing with normalization as in (19)–(22),

$$\mathbf{h}_{\text{proj}} = \text{LayerNorm}(\mathbf{W}_{\text{proj}} \bar{\phi}^{\text{CLS}}(\mathbf{x}) + \mathbf{b}_{\text{proj}}), \quad (19)$$

$$\mathbf{h}_{\text{attn}} = \text{LayerNorm}(\text{MultiHeadAttention}(\mathbf{h}_{\text{proj}}) + \mathbf{h}_{\text{proj}}), \quad (20)$$

$$\mathbf{h}_{\text{mlp}} = \text{LayerNorm}(\text{MLP}(\mathbf{h}_{\text{attn}}) + \mathbf{h}_{\text{attn}}), \quad (21)$$

$$\mathbf{z}(\mathbf{x}) = \mathbf{W}_{\text{head}} \mathbf{h}_{\text{mlp}} + \mathbf{b}_{\text{head}}. \quad (22)$$

The CNN gating network uses FC1 activations because they provide compact mid-level features, and a lightweight MLP can map them into expert weights efficiently. The Mamba gating network leverages state-space projection with multi-head attention to capture sequential structure and contextual dependencies in pre-classification features, which improves adaptive weighting. The ViT gating network applies transformer-style normalization and attention on the CLS token, enabling the gating module to model global context and assign expert weights in a more input-adaptive and context-aware manner.

2) *Gating Network Training Configuration*: Learned gating networks were trained with AdamW (learning rate 0.001, weight decay 10^{-4}), early stopping (patience 15), and gradient clipping (max norm 1.0). The loss function is given in (23),

$$\mathcal{L}_{\text{gating}} = \mathcal{L}_{\text{BCE}}(\hat{y}_{\text{ensemble}}, y) + \lambda \mathcal{L}_{\text{entropy}}(\mathbf{g}), \quad (23)$$

with $\lambda = 0.01$.

Final gating weights for learned networks are normalized as in (24)–(25),

$$g_k^{\text{learned-softmax}}(\mathbf{x}) = \frac{\exp(z_k(\mathbf{x}))}{\sum_{j=1}^K \exp(z_j(\mathbf{x}))}, \quad (24)$$

$$g_k^{\text{learned-sigmoid}}(\mathbf{x}) = \frac{\sigma(z_k(\mathbf{x}))}{\sum_{j=1}^K \sigma(z_j(\mathbf{x}))}. \quad (25)$$

C. End-to-End MoE Training Details

1) *Hierarchical Weight Normalization*: Expert weights are normalized at lobe and lung levels as in (26)–(27),

$$\mathbf{w}_{\text{lobes}} = \frac{\mathbf{w}_{1:5}}{\sum_{i=1}^5 w_i}, \quad (26)$$

$$\mathbf{w}_{\text{lungs}} = \frac{\mathbf{w}_{6:7}}{\sum_{i=6}^7 w_i}. \quad (27)$$

2) *Multi-Component Loss Function*: To ensure that the end-to-end model learns not only accurate predictions but also balanced and interpretable expert contributions, we design a multi-component loss function that combines standard classification loss with additional regularizers for entropy behavior, weight distribution, and expert diversity. The complete loss includes four components in (28)–(31),

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)], \quad (28)$$

$$\mathcal{L}_{\text{entropy}} = -\sum_{k=1}^K w_k \log(w_k + \epsilon), \quad (29)$$

$$\mathcal{L}_{\text{weight}} = \|\mathbf{w} - \frac{1}{K} \mathbf{1}\|_2^2, \quad (30)$$

$$\mathcal{L}_{\text{diversity}} = \frac{1}{K^2} \sum_{i=1}^K \sum_{j \neq i}^K |\cos(\phi_i, \phi_j)|. \quad (31)$$

REFERENCES

- [1] U. Bagci, J. K. Udupa, and L. Bai, “The role of intensity standardization in medical image registration,” *Pattern Recognition Letters*, vol. 31, no. 4, pp. 315–323, 2010.
- [2] A. Jaus, C. Seibold, S. Reiß, L. Heine, A. Schily, M. Kim, F. H. Bahnsen, K. Herrmann, R. Stiefelhagen, and J. Kleesiek, “Anatomy-guided pathology segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 3–13.
- [3] J. Wu, M. Chen, X. Ke, T. Xun, X. Jiang, H. Zhou, L. Shao, and Y. Kong, “Learning heterogeneous tissues with mixture of experts for gigapixel whole slide images,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5144–5153.
- [4] X. Mei, Z. Liu, A. Singh, M. Lange, P. Boddu, J. Q. Gong, J. Lee, C. DeMarco, C. Cao, S. Platt *et al.*, “Interstitial lung disease diagnosis and prognosis using an ai system integrating longitudinal data,” *Nature communications*, vol. 14, no. 1, p. 2272, 2023.
- [5] S. Soffer, A. S. Morgenthau, O. Shimon, Y. Barash, E. Konen, B. S. Glicksberg, and E. Klang, “Artificial intelligence for interstitial lung disease analysis on chest computed tomography: a systematic review,” *Academic Radiology*, vol. 29, pp. S226–S235, 2022.
- [6] S. M. Humphries, D. Thieke, D. Baraghoshi, M. J. Strand, J. J. Swigris, K. J. Chae, H. J. Hwang, A. S. Oh, K. R. Flaherty, A. Adegunsoye *et al.*, “Deep learning classification of usual interstitial pneumonia predicts outcomes,” *American Journal of Respiratory and Critical Care Medicine*, vol. 209, no. 9, pp. 1121–1131, 2024.
- [7] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, “Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images,” in *International MICCAI brainlesion workshop*. Springer, 2021, pp. 272–284.
- [8] K. O’shea and R. Nash, “An introduction to convolutional neural networks,” *arXiv preprint arXiv:1511.08458*, 2015.
- [9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [10] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” *arXiv preprint arXiv:2312.00752*, 2023.
- [11] S. Mu and S. Lin, “A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications,” *arXiv preprint arXiv:2503.07137*, 2025.
- [12] D. Dai, C. Deng, C. Zhao, R. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu *et al.*, “Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models,” *arXiv preprint arXiv:2401.06066*, 2024.
- [13] P. Jin, B. Zhu, L. Yuan, and S. Yan, “Moe++: Accelerating mixture-of-experts methods with zero-computation experts,” *arXiv preprint arXiv:2410.07348*, 2024.
- [14] Y. Lai, X. Liu, Y. Cheng, S. Liu, Y. Wu, W. Zheng *et al.*, “Transformer based multiple superpixel-instance learning for weakly supervised segmenting lesions of interstitial lung disease,” *Expert Systems with Applications*, vol. 253, p. 124270, 2024.
- [15] B. Lin, Z. Tang, Y. Ye, J. Cui, B. Zhu, P. Jin, J. Zhang, M. Ning, and L. Yuan, “Moe-llava: Mixture of experts for large vision-language models,” *arXiv preprint arXiv:2401.15947*, 2024.
- [16] Y. Jiang and Y. Shen, “M4oe: A foundation model for medical multimodal image segmentation with mixture of experts,” in *international conference on medical image computing and computer-assisted intervention*. Springer, 2024, pp. 621–631.
- [17] R. J. Gillies, P. E. Kinahan, and H. Hricak, “Radiomics: images are more than pictures, they are data,” *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.
- [18] R. De Robertis, M. Todesco, D. Autelitano, F. Spoto, and M. D’Onofrio, “The role of radiomics in hepato-bilio-pancreatic surgery: a literature review,” *Artificial Intelligence Surgery*, vol. 3, no. 3, pp. 166–179, 2023.
- [19] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
- [20] Z. Xu, M. Gao, G. Z. Papadakis, B. Luna, S. Jain, D. J. Mollura, and U. Bagci, “Joint solution for pet image segmentation, denoising, and partial volume correction,” *Medical image analysis*, vol. 46, pp. 229–243, 2018.
- [21] A. V. Dalca, J. Guttag, and M. R. Sabuncu, “Anatomical priors in convolutional networks for unsupervised biomedical segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9290–9299.
- [22] N. Khosravan, A. Mortazi, M. Wallace, and U. Bagci, “Pan: Projective adversarial network for medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 68–76.
- [23] U. Bagci, X. Chen, and J. K. Udupa, “Hierarchical scale-based multi-object recognition of 3-d anatomical structures,” *IEEE Transactions on Medical Imaging*, vol. 31, no. 3, pp. 777–789, 2011.
- [24] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghahfarooi, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” *Medical image analysis*, vol. 42, pp. 60–88, 2017.
- [25] Z. Zhang, L. Yao, B. Wang, D. Jha, G. Durak, E. Keles, A. Medetalibeyoglu, and U. Bagci, “Diffboost: Enhancing medical image segmentation via text-guided diffusion model,” *IEEE Transactions on Medical Imaging*, 2024.
- [26] M. Puttagunta and S. Ravi, “Medical image analysis based on deep learning approach,” *Multimedia tools and applications*, vol. 80, no. 16, pp. 24 365–24 398, 2021.
- [27] G. Durak, K. Akin, O. Cetin, E. Uysal, H. E. Aktas, U. Durak, A. Y. Karkas, N. Senkal, H. Savas, A. Tunaci *et al.*, “Radiologic and clinical correlates of long-term post-covid-19 pulmonary sequelae,” *Journal of Clinical Medicine*, vol. 14, no. 14, p. 4874, 2025.

- [28] X. Liu, K. Gao, B. Liu, C. Pan, K. Liang, L. Yan, J. Ma, F. He, S. Zhang, S. Pan *et al.*, “Advances in deep learning-based medical image analysis,” *Health Data Science*, vol. 2021, p. 8786793, 2021.
- [29] A. J. Larrazabal, C. Martinez, and E. Ferrante, “Anatomical priors for image segmentation via post-processing with denoising autoencoders,” in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer, 2019, pp. 585–593.
- [30] D. D. Pham, G. Dovletov, S. Warwas, S. Landgraeber, M. Jäger, and J. Pauli, “Deep learning with anatomical priors: imitating enhanced autoencoders in latent space for improved pelvic bone segmentation in mri,” in *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, 2019, pp. 1166–1169.
- [31] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, and S. Yu, “A survey on incorporating domain knowledge into deep learning for medical image analysis,” *Medical Image Analysis*, vol. 69, p. 101985, 2021.
- [32] K. Wang, Z. Li, H. Wang, S. Liu, M. Pan, M. Wang, S. Wang, and Z. Song, “Improving brain tumor segmentation with anatomical prior-informed pre-training,” *Frontiers in Medicine*, vol. 10, p. 1211800, 2023.
- [33] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, “Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem,” *European radiology experimental*, vol. 4, pp. 1–13, 2020.
- [34] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, “Computational radiomics system to decode the radiographic phenotype,” *Cancer research*, vol. 77, no. 21, pp. e104–e107, 2017.
- [35] H. Wang, Z. Jiang, Y. You, Y. Han, G. Liu, J. Srinivasa, R. Kompella, Z. Wang *et al.*, “Graph mixture of experts: Learning on large-scale graphs with explicit diversity modeling,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 50 825–50 837, 2023.
- [36] S. Chopra, G. Sanchez-Rodriguez, L. Mao, A. J. Feola, J. Li, and Z. Kira, “Medmoe: Modality-specialized mixture of experts for medical vision-language understanding,” *arXiv preprint arXiv:2506.08356*, 2025.