

Hierarchical MoE for Multi-Modal ILD Diagnosis

Alec K. Peltekian¹[0000-0001-9082-4000], Gorkem Durak²[0000-0002-1608-1955],
Halil Ertugrul Aktas²[0009-0009-9491-8122], Carrie Lynn
Richardson³[0000-0002-2541-175X], Mary Carns⁴[0000-0002-5063-156X], Kathleen
Aren³[0000-0003-4770-2643], GR Scott Budinger^{4,5}[0000-0002-3114-5208], Anthony
J. Esposito^{4,5}[0000-0002-8636-0845], Alexander Misharin^{4,5}[0000-0003-2879-3789],
Alok Nidhi Choudhary^{1,6}[0000-0001-8152-6319], Ankit
Agrawal⁶[0000-0002-5519-0302], and Ulas Bagci²[0000-0001-7379-6829] *

¹ Department of Computer Science, Northwestern University McCormick School of Engineering and Applied Science, Chicago, IL, USA

² Department of Radiology, Machine and Hybrid Intelligence Laboratory,
Northwestern University Feinberg School of Medicine, Chicago, IL, USA
ulas.bagci@northwestern.edu

³ Division of Rheumatology, Northwestern University Feinberg School of Medicine,
Chicago, IL, USA

⁴ Division of Pulmonary and Critical Care, Northwestern University Feinberg School
of Medicine, Chicago, IL, USA

⁵ Simpson Querrey Lung Institute for Translational Science, Northwestern University
Feinberg School of Medicine, Chicago, IL, USA

⁶ Department of Electrical and Computer Engineering, Northwestern University
McCormick School of Engineering and Applied Science, Chicago, IL, USA

Abstract. Mixture-of-experts (MoE) models combine specialized predictors under learned routing, offering a principled mechanism for leveraging heterogeneity in medical data. We present a hierarchical multi-modal MoE for interstitial lung disease (ILD) classification that integrates a frozen, pre-trained imaging expert with structured electronic health records (EHR) via two-stage gating. A modality-level gate assigns patient-specific weights to imaging and EHR predictions, while a sub-gating module decomposes the EHR branch into clinically defined feature groups with learned, group-specific contributions. This design preserves stable imaging representations while enabling input-dependent clinical weighting and explicit EHR specialization. Under strict patient-level cross-validation, the model achieved the highest mean AUC among the evaluated methods (0.8750 ± 0.0443), compared with 0.8646 for imaging-only REN and 0.7685 for SwinUNETR. The framework extends interpretability across anatomical regions, imaging-EHR utilization, and clinically defined EHR feature groups.

Keywords: Mixture-of-experts · hierarchical architectures · multimodal learning · radiomics · anatomical priors · interstitial lung disease · medical imaging · deep learning

* Corresponding author.

1 Introduction

Mixture-of-experts (MoE) architectures perform conditional computation by routing inputs to specialized sub-networks [20, 6, 16]. Most existing MoE designs assume homogeneous expert capacity and domain-agnostic routing [19, 3], which is misaligned with medical imaging, where anatomical structure and region-specific pathology impose strong constraints on expert specialization and require interpretable, domain-aware modeling [13, 4].

Pulmonary imaging illustrates this mismatch. Interstitial lung disease (ILD) exhibits heterogeneous, regionally distributed involvement; clinical interpretation of high-resolution CT is inherently regional, with patterns such as basal and subpleural predominance [18, 5]. Despite this, most deep-learning approaches process the lung as a single global volume, diluting regional signals and limiting interpretability [10]. Anatomically structured MoE models have improved ILD classification by promoting region-aware specialization [17], yet remain imaging-only and do not incorporate the structured clinical information that is central to ILD diagnosis in practice [1].

Integrating imaging and structured EHR can be formally viewed as a heterogeneous conditional risk minimization problem in which the optimal predictive function depends on patient-specific modality relevance. Let X_I denote imaging features and X_E structured clinical variables. The Bayes-optimal decision rule for ILD diagnosis is generally not separable across modalities and may vary across subpopulations. Fixed fusion strategies implicitly assume uniform modality utility, resulting in suboptimal bias when modality informativeness is patient-dependent. We therefore formulate multimodal ILD classification as a structured conditional computation problem, where routing mechanisms approximate a patient-specific decomposition of the predictive function:

$$f(X_I, X_E) = \sum_m g_m(X_I, X_E) f_m(X_I, X_E), \quad (1)$$

where gating functions g_m dynamically allocate representational capacity across modality- and domain-specialized experts. Under this perspective, effective multimodal modeling requires not only feature fusion but principled routing that reflects anatomical and clinical structure.

The two modalities encode complementary but heterogeneous information whose relative diagnostic value varies across patients, comorbidities, and disease stages [9, 15]. Prior multimodal approaches have explored feature concatenation, attention-based fusion, and cross-modal transformers [12, 11], yet typically rely on dense interaction or fixed fusion strategies without explicit, structured routing across modalities or clinically defined feature groups. In some cases, radiologic findings alone suffice to exclude disease; in others, subtle impairments in pulmonary function or clinical phenotype provide the decisive signal. Effective multimodal models must therefore adaptively weight information across modalities and clinical domains while maintaining the interpretability required for clinical deployment.

We introduce a hierarchical multimodal MoE architecture for ILD classification with two-level gating (Fig. 1). A *modality-level gate* assigns patient-specific weights to an imaging expert and an EHR branch, enabling adaptive reliance on each modality. Within the EHR branch, a *sub-gating module* decomposes structured clinical variables into clinically defined feature groups and learns group-specific contributions, providing structured, input-dependent fusion while preserving interpretability. We evaluate the model under patient-level cross-validation and achieve the highest mean AUC among the evaluated methods, together with case-level summaries of modality and clinical-feature usage derived from learned gate activations.

2 Methods

Dataset and Preprocessing. We used a retrospective cohort from the Northwestern Scleroderma Registry: 597 patients with 1,898 longitudinal chest CT scans (2001–2023). The cohort comprised 489 (81.9%) female patients (mean age 63.7 ± 12.7 yr). Disease subtypes included limited cutaneous SSc (47.6%), diffuse cutaneous SSc (41.0%), SSc sine scleroderma (4.7%), and other (6.7%). ILD was confirmed in 365 patients (61.1%), forming the positive class.

CT volumes were resampled to isotropic 1 mm spacing, segmented into five anatomical lung lobes (LUL, LLL, RUL, RML, RLL) using LungMask R231 [8], intensity-windowed to $[-175, 250]$ HU, and resized to $96 \times 96 \times 96$ voxels. The study was approved by the Northwestern IRB with informed consent from all participants.

Radiomics-Guided Lobe Importance. Following anatomically informed MoE work [17], we estimate lobe-level diagnostic importance via radiomics. For each cross-validation fold, 107 PyRadiomics features [21] are extracted per lobe, and lobe-specific XGBoost classifiers [2] are trained on the training split. Validation AUCs are converted into fixed attention weights:

$$w_k = \begin{cases} 0.1 + (\text{AUC}_k - 0.5) \cdot 3.8, & \text{AUC}_k \geq 0.5 \\ 0.1, & \text{otherwise} \end{cases} \quad (2)$$

This mapping assigns a minimum nonzero contribution to lobes with validation AUC below chance and increases the weight linearly for more predictive lobes. The weights are computed once per fold and held fixed during multimodal training. Alternative mappings were not systematically evaluated.

Lobe-Aware Imaging Expert. The imaging branch consists of five lobe-specific experts (LUL, LLL, RUL, RML, RLL), each implemented using a 3D SwinUNETR backbone [7]. Lobe masks restrict feature extraction to anatomically meaningful regions. Each lobe expert produces a representation $\phi_{\text{img},k}$, and these are aggregated using the fixed radiomics-derived lobe importance weights from Eq. (2).

This fold-specific weighting can be interpreted as a data-dependent prior over expert relevance, introducing a structured inductive bias into the routing

process. By decoupling anatomical importance estimation from gradient-based optimization, we constrain the imaging MoE to regionally consistent solutions while preserving specialization. Formally, this implements a constrained optimization:

$$\min_{\theta} \mathcal{L}(f_{\theta}) \quad \text{s.t.} \quad w_k = h(\text{AUC}_k), \quad (3)$$

where $h(\cdot)$ is the lower-bounded affine mapping defined in Eq. (2). The resulting weights are normalized before expert aggregation:

$$\tilde{w}_k = \frac{w_k}{\sum_{j=1}^5 w_j}, \quad \phi_{\text{img}} = \sum_{k=1}^5 \tilde{w}_k \phi_{\text{img},k}. \quad (4)$$

Global average pooling and a projection head yield a 48-dimensional imaging representation $\phi_{\text{img}} \in \mathbb{R}^{48}$. The imaging expert is trained independently on CT data and *frozen* during multimodal training, preserving lobe-specific specialization while the EHR branch and gating networks are optimized.

EHR Processing and Feature Grouping. Structured EHR variables were linked separately to each CT examination. For each variable, the most recent measurement recorded on or before the CT date was selected. When no prior measurement was available, the nearest subsequent measurement within a 30-day grace window was permitted. Values unavailable within this interval were zero-filled before standardization.

EHR features were standardized independently within each cross-validation fold using feature-wise means and standard deviations computed exclusively from the training partition. The same transformation was applied unchanged to the validation and test partitions. Explicit missingness indicators were not used.

To enable hierarchical specialization, EHR variables were partitioned into clinically coherent groups. We evaluated two configurations: **(i) Full hierarchy (7 groups; 69 variables total)**. EHR variables were partitioned into seven clinically coherent groups: Complete Blood Count (21 variables), Pulmonary Function Tests (3 variables), Serum Chemistries (14 variables), Laboratory and Functional Markers (4 variables), Vital Signs (7 variables), Static Patient Demographics (14 variables), and Disease-Specific Demographics (6 variables).

(ii) Selective hierarchy (3 groups; 12 variables total). The selective configuration restricts the EHR branch to three clinically proximal groups: Pulmonary Function Tests (FVC, FEV1, DLCO), Disease-Specific Demographics (systemic sclerosis subtype indicators [lcSSc, dcSSc, SSS] and autoantibody status including Scl-70, anticentromere antibody, and RNA polymerase III), and Key Biomarkers (peak creatinine, modified Rodnan skin score, and B-type natriuretic peptide). This selective grouping concentrates routing capacity on dynamic and mechanistically relevant predictors while excluding broader static demographic variables.

Hierarchical EHR Expert. Grouping EHR variables into clinically coherent subsets imposes a structured sparsity prior over multimodal interactions. Without grouping, dense fusion induces high-variance cross-modal entanglement, especially when weakly informative or static variables are present. Let

$X_E = \{X_1, \dots, X_G\}$ denote grouped features. The hierarchical MoE approximates:

$$f_E(X_E) = \sum_{g=1}^G \pi_g(X_E) f_g(X_g), \quad (5)$$

where f_g are group-specific sub-experts and π_g are input-dependent routing weights. This decomposition reduces the effective interaction dimensionality from $\mathcal{O}(|X_E|^2)$ to $\mathcal{O}(G)$, controlling variance while preserving conditional flexibility.

Let $\mathbf{x}_g$ denote the features belonging to group g . Each group is processed by a dedicated sub-expert with residual projection:

$$\phi_g = \text{MLP}_g(\mathbf{x}_g) + \text{Proj}_g(\mathbf{x}_g), \quad \phi_g \in \mathbb{R}^{24}. \quad (6)$$

Sub-expert outputs are aggregated via a learned gating network:

$$\mathbf{g}_{\text{sub}} = \text{Softmax}(\text{Gate}_{\text{sub}}([\phi_1; \dots; \phi_G])). \quad (7)$$

The final EHR representation combines gated and direct information through a dual-fusion strategy:

$$\phi_{\text{ehr}} = \text{Fusion}([\phi_1; \dots; \phi_G]) + \text{Proj}\left(\sum_{g=1}^G g_{\text{sub},g} \phi_g\right), \quad (8)$$

yielding $\phi_{\text{ehr}} \in \mathbb{R}^{48}$.

Modality-Level Gating and Classification. A modality gate adaptively balances imaging and EHR contributions:

$$\mathbf{g}_{\text{mod}} = \text{Softmax}(\text{Gate}_{\text{mod}}([\phi_{\text{img}}; \phi_{\text{ehr}}])) \in \mathbb{R}^2. \quad (9)$$

The fused representation $\phi_{\text{final}} = g_{\text{mod},1} \phi_{\text{img}} + g_{\text{mod},2} \phi_{\text{ehr}}$ is passed to a lightweight MLP classifier for binary ILD prediction.

Training Protocol. Training proceeds in three stages: (1) radiomics-based lobe importance estimation, (2) imaging expert training with lobe-aware attention, and (3) multimodal training with the imaging expert frozen while EHR sub-experts, gating networks, and the classifier are optimized jointly. All models use AdamW (lr = 4×10^{-4} , weight decay 10^{-5}) [14], batch size 4, early stopping (patience 10), and inverse-frequency class weighting. Evaluation uses strict patient-level 5-fold cross-validation, with all longitudinal CT scans from each patient retained in the same fold to prevent patient-level leakage. AUC is reported as mean $\pm$ SD across folds. Statistical comparisons used two-sided paired t -tests on fold-level AUCs obtained from identical patient-level partitions; the selective hierarchical MoE was additionally compared directly with imaging-only REN. Data augmentation includes random rotation ($\pm 15^\circ$), scaling (0.9–1.1), and Gaussian noise. All models were implemented in PyTorch/MONAI on NVIDIA A100 GPUs.

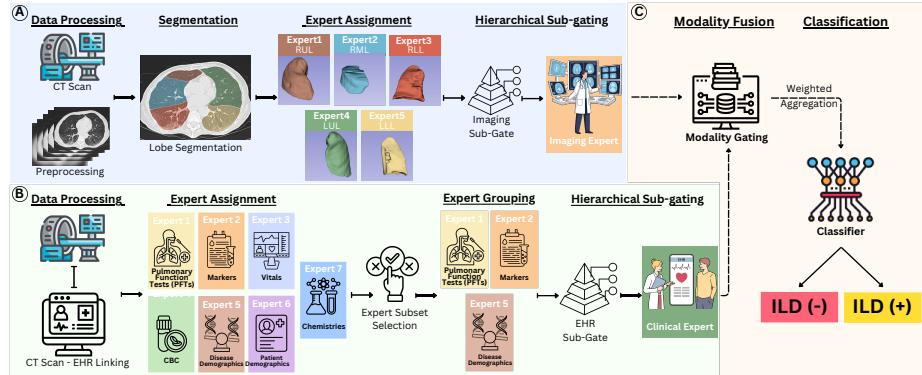


Fig. 1. Overview of the proposed hierarchical multimodal MoE framework for ILD diagnosis. **(A)** CT pipeline: volumes are preprocessed and segmented into five lung lobes; five lobe-specific imaging experts operate on lobe-masked volumes; fixed radiomics-derived lobe-importance weights provide a fold-specific prior for aggregating expert representations. **(B)** EHR pipeline: structured variables are partitioned into clinically defined feature groups and processed by group-specific sub-experts; an EHR sub-gate aggregates sub-expert outputs into a single clinical representation. We additionally evaluate a *selective* EHR configuration via expert-subset selection (restricting to a targeted subset of feature groups) prior to EHR sub-gating. **(C)** Multimodal fusion: a modality gate adaptively weights imaging and clinical representations, followed by a classifier for ILD prediction.

3 Results

Table 1 summarizes performance across all methods. The selective hierarchical MoE achieved the highest mean AUC (0.8750 ± 0.0443), compared with 0.8646 ± 0.0467 for imaging-only REN, corresponding to an absolute AUC increase of 0.0104. However, the paired fold-level difference was not statistically significant ($t(4) = 0.265$, $p = 0.804$), so the numerical improvement requires confirmation in a larger evaluation. The selective model significantly outperformed SwinUNETR ($p < 0.001$). Notably, the highest mean AUC was achieved with the imaging expert frozen while only the EHR branch, gating networks, and final classifier were optimized.

Effect of Hierarchical EHR Grouping. The selective hierarchy achieved a higher mean AUC than the full hierarchy (0.8750 vs. 0.8496). A plausible explanation is that the full configuration distributes routing capacity across 69 variables and seven groups, including static or potentially redundant variables. The selective hierarchy instead concentrates routing on pulmonary function, disease phenotype and serology, and key biomarkers that are more proximal to ILD status. This interpretation is consistent with the concentration of selective sub-gate weights on the PFT expert, although it does not establish that the excluded variables are intrinsically uninformative.

Robustness, longitudinal stability, and fusion baselines.

Table 1. Performance comparison using 5-fold patient-level cross-validation. Values are reported as mean $\pm$ SD across folds. Table p -values are from two-sided paired t -tests against SwinUNETR. The direct comparison between the selective hierarchical MoE and imaging-only REN is reported in the text.

Method	AUC $\pm$ SD [95% CI]	Change vs. SwinUNETR	p -value
<i>Imaging-Only Baselines</i>			
SwinUNETR (Baseline)	0.7685 ± 0.0759 [0.674, 0.863]	–	–
CNN (Baseline)	0.7584 ± 0.0919 [0.667, 0.850]	–1.3%	–
Mamba (Baseline)	0.6775 ± 0.0454 [0.632, 0.723]	–11.8%	–
ViT (Baseline)	0.6535 ± 0.0356 [0.618, 0.689]	–15.0%	–
<i>MoE Anatomically-Guided Imaging-Only (REN)</i>			
Radiomics-Guided REN	0.8646 ± 0.0467 [0.806, 0.923]	+12.5%	0.031
<i>Multi-Modal Approaches</i>			
Hierarchical MoE (Full EHR)	0.8496 ± 0.0449 [0.794, 0.905]	+10.6%	0.004
Hierarchical MoE (Selective EHR)	0.8750 ± 0.0443 [0.820, 0.930]	+13.9%	< 0.001

Temporal linkage sensitivity: Varying the forward grace window used for EHR–CT matching when no prior measurement existed (0, 7, 14, 30, 60, and 90 days) produced negligible changes in performance (AUC 0.8746 ± 0.0443 , 0.8746 ± 0.0443 , 0.8749 ± 0.0442 , 0.8750 ± 0.0443 , 0.8750 ± 0.0443 , and 0.8750 ± 0.0441 , respectively), indicating robustness to reasonable temporal linkage policies.

Longitudinal prediction variance: Among patients with more than one CT scan ($n = 192$), the mean within-patient standard deviation of predicted ILD probability was 0.0812 ± 0.0964 , with a mean within-patient probability range of 0.2004 ± 0.2406 . Variability was slightly lower in ILD-positive patients ($n = 138$; standard deviation 0.0756 ± 0.0922 , range 0.1901 ± 0.2344) than in ILD-negative patients ($n = 54$; standard deviation 0.0953 ± 0.1060 , range 0.2265 ± 0.2562), consistent with higher uncertainty in borderline-negative cases.

Multimodal integration and late-fusion baselines: Under identical patient-level folds, the selective hierarchical MoE achieved 0.8750 ± 0.0443 AUC, compared with 0.8646 ± 0.0467 for imaging-only REN. Among simpler integration strategies, concatenation followed by logistic regression achieved 0.8595 ± 0.0438 , concatenation with an MLP achieved 0.8337 ± 0.0551 , and EHR-only logistic regression achieved 0.8280 ± 0.0517 . These results show that the complete hierarchical model achieved the highest mean AUC among the evaluated fusion strategies, although they do not independently isolate the contributions of the modality gate and EHR sub-gate.

Interpretability of Hierarchical Gating. We analyzed learned gate activations at both the modality and EHR sub-expert levels (Fig. 2). Gate activations represent model routing behavior rather than causal feature importance. A high gate weight indicates greater utilization of an expert but does not establish its independent clinical importance.

Modality-level routing: Imaging experts receive higher weights for ILD-negative predictions, consistent with clear parenchyma being sufficient to exclude disease. EHR experts are weighted more heavily for ILD-positive predictions, particularly

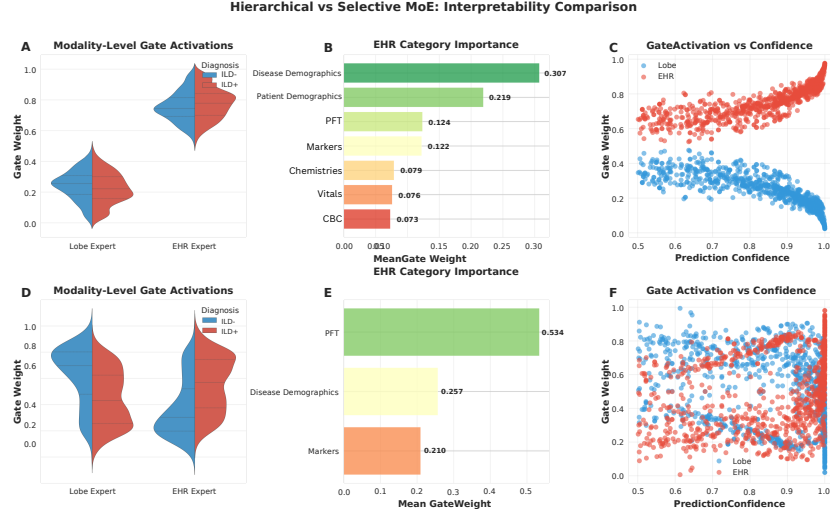


Fig. 2. Interpretability analysis of hierarchical MoE gating behavior. (A) Modality-level gate activations for the full hierarchical EHR configuration, stratified by ILD status. (B) Mean EHR sub-expert gate weights across clinical feature groups for the full hierarchy. (C) Relationship between modality-level gate activations and prediction confidence for the full hierarchy. (D-F) Corresponding analyses for the selective hierarchical EHR configuration.

at higher confidence, mirroring clinical workflows in which confirmation of ILD relies on pulmonary function impairment and disease phenotype.

EHR sub-expert routing: In the full hierarchy, routing is distributed broadly across clinical categories, with increasing EHR dominance at high confidence—a pattern associated with overconfident positive predictions. In the selective hierarchy, routing concentrates on PFT sub-experts, followed by disease demographics and biomarkers, while modality-level gating maintains more balanced imaging–EHR usage across confidence levels. Together, these patterns provide a model-level interpretation of the observed performance difference: restricting EHR integration to clinically proximal groups reduces excessive EHR dominance, preserves complementary imaging evidence, and improves calibration.

Calibration and Error Analysis: Across five folds, the selective hierarchical model achieved an ECE of 0.131 ± 0.038 and a Brier score of 0.135 ± 0.037 , indicating reasonable calibration. Of 960 test samples, 167 (17.4%) were misclassified, with 71 errors (42.5%) at high confidence (≥ 0.8). Misclassified cases exhibited lower EHR reliance (mean gate weight 0.383 vs. 0.527 for correct predictions), while false positives were more strongly imaging-driven. These patterns suggest high-confidence errors are associated with imaging-dominant routing in clinically ambiguous cases, motivating confidence-aware routing in future work.

Modality Contribution Analysis: The selective hierarchical MoE achieves balanced average modality weights (imaging 0.498 ± 0.122 , EHR 0.502 ± 0.122), con-

trasting with fixed or late-fusion strategies that assume static contributions. Learned gates adaptively emphasize imaging or EHR per sample, providing a soft, patient-specific alternative to non-hierarchical fusion.

4 Discussion and Concluding Remarks

We introduced a hierarchical multimodal MoE framework that extends anatomically informed REN with interpretable routing across imaging, EHR, and clinically defined feature groups. The selective configuration achieved the highest mean AUC while providing case-specific summaries of anatomical, modality, and EHR-group utilization. Although the numerical improvement over imaging-only REN was not statistically significant, the results motivate further evaluation of clinically structured multimodal routing in larger and external cohorts.

Limitations and future work. Our cohort originates from a single academic health system and focuses on scleroderma-associated ILD, which may limit generalizability to other etiologies. However, patients were imaged across multiple affiliated hospitals with diverse scanners over two decades, reducing the likelihood that performance reflects a narrow acquisition pipeline. EHR feature grouping relies on clinically motivated rather than learned partitions, and data-driven groupings may capture additional interactions. The framework also depends on automated lung and lobe segmentation, introducing potential error propagation. Missing EHR values were represented without explicit missingness indicators, which may limit the model’s ability to distinguish unavailable measurements from observed values after preprocessing. Furthermore, the current comparisons do not independently isolate the contributions of the modality gate and EHR sub-gate. Future work should evaluate explicit missingness modeling, component-wise gating ablations, patient-level endpoints, multi-institutional external validation, longitudinal progression and subtype-aware prediction, and model compression for deployment. Although developed for systemic sclerosis-associated ILD, the framework may also be applicable to other multimodal diagnostic tasks involving structured imaging and clinical data.

References

1. Chatterjee, S., Perelas, A., Yadav, R., Kirby, D.F., Singh, A.: a multidisciplinary approach to the assessment of patients with systemic sclerosis-associated interstitial lung disease. *Clinical Rheumatology* **42**(3), 653–661 (2023)
2. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. pp. 785–794 (2016)
3. Chen, Z., Deng, Y., Wu, Y., Gu, Q., Li, Y.: Towards understanding the mixture-of-experts layer in deep learning. *Advances in neural information processing systems* **35**, 23049–23062 (2022)
4. Dalca, A.V., Gutttag, J., Sabuncu, M.R.: Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 9290–9299 (2018)

5. Desai, S.R., Veeraraghavan, S., Hansell, D.M., Nikolakopoulou, A., Goh, N.S., Nicholson, A.G., Colby, T.V., Denton, C.P., Black, C.M., Du Bois, R.M., et al.: Ct features of lung disease in patients with systemic sclerosis: comparison with idiopathic pulmonary fibrosis and nonspecific interstitial pneumonia. *Radiology* **232**(2), 560–567 (2004)
6. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* **23**(120), 1–39 (2022)
7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: *International MICCAI brainlesion workshop*. pp. 272–284. Springer (2021)
8. Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., Langs, G.: Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European radiology experimental* **4**, 1–13 (2020)
9. Huang, S.C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.P.: Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine* **3**(1), 136 (2020)
10. Humphries, S.M., Thieke, D., Baraghoshi, D., Strand, M.J., Swigris, J.J., Chae, K.J., Hwang, H.J., Oh, A.S., Flaherty, K.R., Adegunsoye, A., et al.: Deep learning classification of usual interstitial pneumonia predicts outcomes. *American Journal of Respiratory and Critical Care Medicine* **209**(9), 1121–1131 (2024)
11. Kamrul, S.M., Jobair, H.F.M., Bofan, H., Cheng, J.Q., Huanying, G.: Multimodal models in healthcare: Methods, challenges, and future directions for enhanced clinical decision support. *Information* **16**(11), 971 (2025)
12. Li, Y., Daho, M.E.H., Conze, P.H., Zeghlache, R., Le Boité, H., Tadayoni, R., Cochener, B., Lamard, M., Quéllec, G.: A review of deep learning-based information fusion techniques for multimodal medical image classification. *Computers in Biology and Medicine* **177**, 108635 (2024)
13. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. *Medical image analysis* **42**, 60–88 (2017)
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017)
15. Mohsen, F., Ali, H., El Hajj, N., Shah, Z.: Artificial intelligence-based methods for fusion of electronic health records and imaging data. *Scientific Reports* **12**(1), 17981 (2022)
16. Mu, S., Lin, S.: A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv preprint arXiv:2503.07137* (2025)
17. Peltekian, A.K., Aktas, H.E., Durak, G., Grudzinski, K., Bemiss, B.C., Richardson, C., Dematte, J.E., Budinger, G., Esposito, A.J., Misharin, A., et al.: Ren: Anatomically-informed mixture-of-experts for interstitial lung disease diagnosis. *arXiv preprint arXiv:2510.04923* (2025)
18. Raghu, G., Remy-Jardin, M., Myers, J.L., Richeldi, L., Ryerson, C.J., Lederer, D.J., Behr, J., Cottin, V., Danoff, S.K., Morell, F., et al.: Diagnosis of idiopathic pulmonary fibrosis. an official ats/ers/jrs/alat clinical practice guideline. *American journal of respiratory and critical care medicine* **198**(5), e44–e68 (2018)
19. Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., Keysers, D., Houlsby, N.: Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems* **34**, 8583–8595 (2021)

20. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
21. Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.: Computational radiomics system to decode the radiographic phenotype. *Cancer research* **77**(21), e104–e107 (2017)