RA-HMFNet: Region-Aware Hierarchical Multimodal Fusion for Preoperative Two-Year Recurrence Prediction in NSCLC

Anonymous authors
Paper under double-blind review

Abstract

Reliable preoperative estimation of postoperative recurrence risk in operable non-small-cell lung cancer (NSCLC) is difficult because CT, ^{18}F-FDG PET, and clinical variables characterize complementary aspects of tumor phenotype, while the spatially heterogeneous peritumoral environment is often reduced to a single handcrafted region. We propose RA-HMFNet, a region-aware hierarchical multimodal fusion network for preoperative prediction of recurrence within 730 days after surgery. On a shared CT deep feature map, non-overlapping 0–3, 3–6, and 6–9 mm peritumoral masks perform region-aware pooling to retain distance-specific representations. PET and clinical embeddings then condition patient-specific scale attention, after which intratumoral, peritumoral, PET, and clinical representations are fused hierarchically. We evaluate the framework on 107 patients (25 recurrences) from the public NSCLC-Radiogenomics dataset using five repeats of five-fold outer stratified cross-validation with three-fold inner cross-validation; all reported metrics use each patient's mean outer-test out-of-fold probability. RA-HMFNet achieved AUROC 0.6634, AUPRC 0.3861, Brier score 0.1683, and log loss 0.5146. Matched ablations yielded AUROC 0.5810 for NoROI, 0.6478 for a single 0–9 mm peritumoral ROI, 0.6390 for fixed equal three-ring fusion, and 0.6566 for globally learned scale weights. Thus, explicit peritumoral localization produced the largest numerical change, whereas distance-specific decomposition alone did not improve over a single 0–9 mm region; learnable scale aggregation recovered higher discrimination. Attention weights remained near uniform and only moderately repeatable, so they are not interpreted as biological importance. The study provides internally validated descriptive evidence and requires independent external validation.

1 Introduction

Postoperative recurrence remains a central challenge after curative-intent resection of NSCLC. A clinically useful preoperative prediction system must work with information available before surgery and integrate signals distributed across anatomical imaging, metabolic imaging, and patient-level clinical context. CT captures morphology and surrounding lung structure, ^{18}F-FDG PET captures metabolic activity, and clinical variables summarize complementary host and disease context. Treating these sources as interchangeable feature vectors can discard the hierarchy and locality that make them informative.

The peritumoral region is particularly relevant to representation design. Prior CT radiomics studies have shown that features extracted outside the gross tumor boundary can contribute to recurrence or survival prediction (D'Antonoli et al., 2020; Lee et al., 2022). Lee et al. explicitly compared size-variant peritumoral regions for two-year recurrence-free survival (Lee et al., 2022). These studies motivate spatially structured modeling, but most rely on engineered radiomics or a single expanded ROI. For deep representation learning, a more specific question is whether the peritumoral area should merely be localized or further decomposed by distance and fused adaptively.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Multimodal PET/CT prognostic modeling provides a second motivation. PET radiomics and PET/CT fusion have shown predictive value in surgically treated NSCLC cohorts (Kirienko et al., 2018; Amini et al., 2021), while more recent work has examined PET/CT/clinical fusion for postoperative recurrence prediction (Mehri-Kakavand et al., 2025; Park et al., 2023). However, direct concatenation treats modality integration as a single-stage operation and does not explicitly separate local CT spatial aggregation from cross-modal interaction.

We address these issues with RA-HMFNet. The model first introduces an explicit spatial inductive bias by aggregating deep CT features within non-overlapping distance-defined peritumoral masks. PET and clinical context then condition the contribution of the three peritumoral scales, followed by a second fusion stage across intratumoral CT, aggregated peritumoral CT, PET, and clinical representations. The final architecture contains approximately 4.40 million trainable parameters and uses Group Normalization to accommodate the small micro-batches required by 3D imaging (Wu & He, 2018).

Our experimental design is organized around four questions: whether explicit ROI localization matters, whether distance-specific decomposition improves upon a single 0–9 mm ROI, whether globally learnable scale weighting improves upon equal weighting, and whether patient-specific weighting provides an additional numerical gain. These questions are evaluated with repeated nested validation and matched architectural controls so that model-selection decisions remain separated from final performance estimation (Bradshaw et al., 2023).

Our contributions are threefold. First, we introduce feature-map-level ROI-aware pooling for non-overlapping 0–3, 3–6, and 6–9 mm peritumoral regions. Second, we develop PET/clinical-conditioned patient-specific scale aggregation followed by hierarchical multimodal fusion. Third, we use a matched mechanistic ablation sequence—NoROI, P09, ROI Equal, ROI Global, and the full patient-specific model—to distinguish spatial localization, distance decomposition, globally learned weighting, and patient-specific weighting.

2 Preliminaries

We introduce the basic operations underlying peritumoral representation and multimodal aggregation. These formulations also define the controls used later to isolate the effects of spatial localization and adaptive scale fusion.

2.1 Distance-defined peritumoral regions

Let M_{IT} denote the binary intratumoral mask and $\mathcal{D}_r(M_{\mathrm{IT}})$ its morphological dilation by physical distance r. A non-overlapping peritumoral ring between distances a and b from the tumor boundary can be expressed as

$$M_{a:b} = \mathcal{D}_b(M_{\mathrm{IT}}) \setminus \mathcal{D}_a(M_{\mathrm{IT}}), \qquad 0 \le a < b. \tag{1}$$

We use $M_{0:3}$, $M_{3:6}$, and $M_{6:9}$, denoted as P03, P36, and P69. Their union defines the single 0–9 mm peritumoral control,

$$M_{\mathrm{P09}} = M_{0:3} \cup M_{3:6} \cup M_{6:9}. \tag{2}$$

Prior NSCLC studies commonly extract handcrafted radiomic descriptors from intratumoral and expanded peritumoral ROIs (D'Antonoli et al., 2020; Lee et al., 2022). Such a representation can be abstractly written as $x_r = \phi(I, M_r)$, where $\phi(\cdot)$ is a radiomic feature extractor. This summarizes an ROI independently and does not explicitly preserve its correspondence within a learned feature map.

2.2 Global and region-aware deep pooling

Given a deep feature map $F \in \mathbb{R}^{C\times D\times H\times W}$, conventional global average pooling summarizes all spatial locations:

$$z_{\mathrm{GAP}} = \frac{1}{|\Omega|} \sum_{v\in\Omega} F(v), \tag{3}$$

where Ω denotes the full spatial support of the feature map. This operation is compact but does not distinguish features originating from different peritumoral distances.

A region-specific representation can instead be obtained by masked feature-map pooling:

$$\mathrm{Pool}(F, M) = \frac{\sum_v \widetilde{M}(v) F(v)}{\sum_v \widetilde{M}(v) + \epsilon}, \tag{4}$$

where $\widetilde{M}$ is the ROI mask mapped to the spatial resolution of F. This preserves a shared feature extractor while allowing different spatial regions to produce distinct representations. For non-overlapping regions, pooled representations over their union can be decomposed into a geometry-weighted combination of the corresponding region-specific representations under the same feature map and pooling operator. This decomposition clarifies what spatial information is collapsed before task-dependent fusion; a derivation is provided in Appendix C.

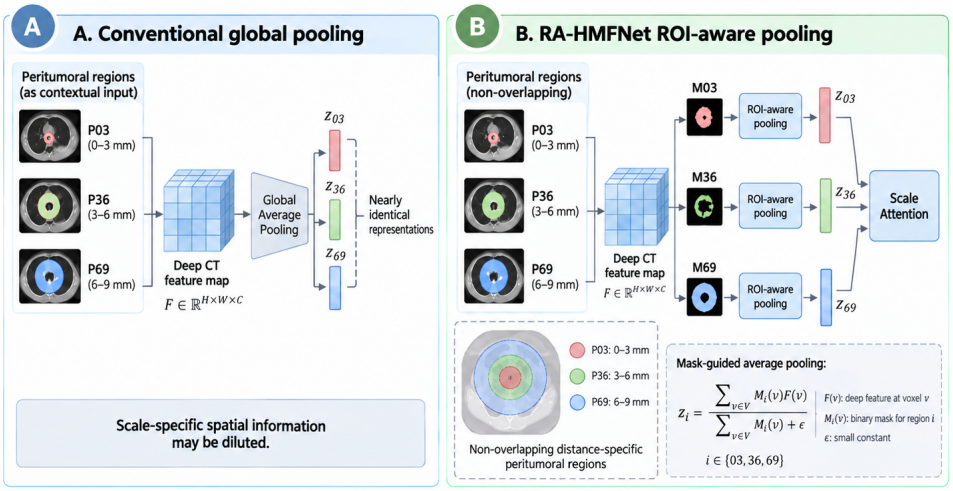


Figure 1: ROI-aware feature-map pooling. Non-overlapping distance-defined masks are mapped to the deep feature grid to aggregate P03, P36, and P69 representations while sharing the peritumoral encoder.

2.3 Multimodal and attention-based aggregation

A common multimodal strategy is direct feature concatenation. For CT, PET, and clinical representations,

$$z_{\mathrm{cat}} = [z_{\mathrm{CT}}; z_{\mathrm{PET}}; z_C], \qquad \hat{p} = \sigma\left(w^\top z_{\mathrm{cat}} + b\right), \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid function. PET/CT studies have used feature- and multimodal-level fusion to integrate anatomical, metabolic, and clinical information (Kirienko et al., 2018; Amini et al., 2021; Mehri-Kakavand et al., 2025; Park et al., 2023).

More generally, attention-based aggregation assigns normalized weights to representations $\{z_i\}_{i=1}^K$:

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^K \exp(s_j)}, \qquad z = \sum_{i=1}^K \alpha_i z_i. \tag{6}$$

The relevant question for peritumoral modeling is therefore not only whether multiple distance-specific representations should be constructed, but also how their contributions should be determined. RA-HMFNet addresses this by conditioning scale scores jointly on PET and clinical context.

3 Method

3.1 Problem definition and cohort

We formulate the task as binary preoperative prediction of postoperative recurrence within 730 days. For patient n, the input is $X_n = \{X_{\mathrm{CT},n}, X_{\mathrm{PET},n}, x_{C,n}\}$ with label $y_n \in \{0, 1\}$, where $y_n = 1$ denotes documented recurrence within 730 days after surgery. The model predicts $p_n = P(y_n = 1 \mid X_n)$. The analysis uses the public NSCLC-Radiogenomics resource (Bakr et al., 2018). After applying the frozen multimodal eligibility criteria, the final cohort contains 107 patients, including 25 recurrences within 730 days and 82 negatives. No postoperative or outcome-derived variables are used as model inputs. Detailed endpoint and cohort criteria are provided in Appendix B.

3.2 Imaging and clinical preprocessing

CT volumes and tumor masks were transformed to $1 \times 1 \times 1$ mm^3 geometry. We retained the intratumoral mask and generated three non-overlapping peritumoral rings: P03 (0–3 mm), P36 (3–6 mm), and P69 (6–9 mm); their union defines the P09 control. CT and registered PET volumes were cropped to a common $128 \times 128 \times 128$ spatial window. CT intensities were clipped to $[-1000, 400]$ HU and normalized to $[-1, 1]$, while PET was scaled after quantitative preprocessing. The clinical branch uses 14 preoperative variables. All preprocessing steps that require fitted parameters are estimated only from the corresponding training partition. Additional PET-registration, clinical-variable, and endpoint details are provided in Appendix B.

3.3 RA-HMFNet architecture

3.3.1 Region-aware deep feature aggregation

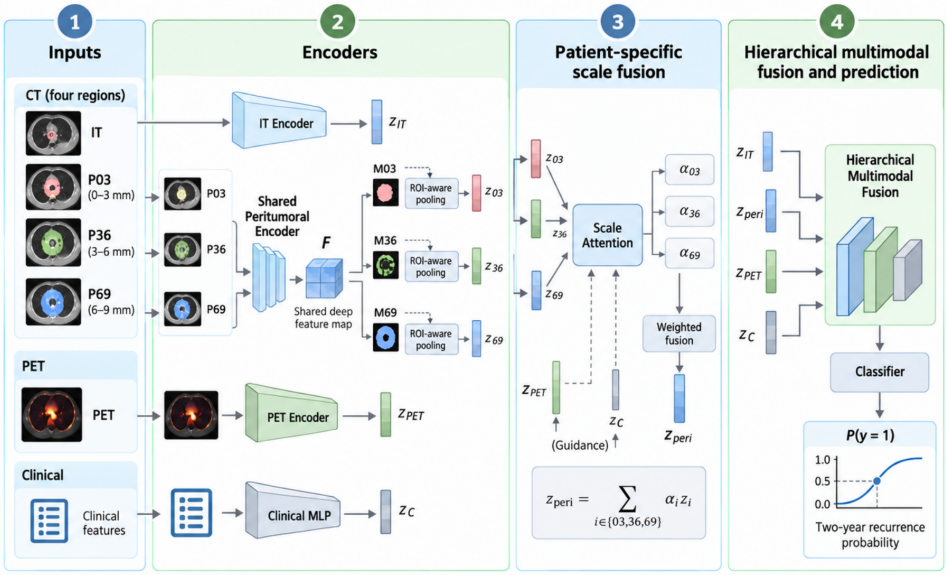


Figure 2: RA-HMFNet architecture. Distance-specific P03/P36/P69 features are pooled from a shared CT feature map, conditioned by PET and clinical context, aggregated into a peritumoral representation, and fused with IT, PET, and clinical representations.

RA-HMFNet contains an intratumoral CT encoder, a shared peritumoral CT encoder, a PET encoder, and a clinical representation branch. The imaging encoders produce 128-dimensional embeddings and use Group Normalization. The final network contains approximately 4.40 million trainable parameters. The architecture is intentionally organized as a

sequence of constrained representation steps rather than a single joint fusion block. Intratumoral CT and peritumoral CT are encoded separately because they correspond to distinct anatomical supports, while the three peritumoral scales share one encoder so that scale comparisons are made in a common feature space. PET and clinical branches are retained as independent sources of context instead of being absorbed into the CT representation. This design keeps the origin of each representation explicit before fusion and makes the subsequent matched ablations easier to interpret.

Let $F \in \mathbb{R}^{C\times D\times H\times W}$ denote the shared peritumoral CT feature map. Each distance-defined mask M_i, $i \in \{03, 36, 69\}$, is mapped to feature-map resolution. Using the operator in Eq. (4),

$$z_i = \mathrm{Pool}(F, M_i), \qquad i \in \{03, 36, 69\}. \tag{7}$$

Unlike conventional global pooling in Eq. (3), this operation retains explicit correspondence between learned CT features and distance-defined peritumoral regions while sharing one peritumoral encoder across scales. Because all three regional vectors are extracted from the same deep feature map, differences among z_{03}, z_{36}, and z_{69} arise from spatial support rather than from independently parameterized encoders. This keeps the comparison among scales controlled and avoids confounding distance-specific representation with scale-specific network capacity. It also allows the P09 control to isolate the effect of preserving separate regional readouts versus pooling the same peritumoral neighborhood as a single region. A further advantage of this construction is that the scale-specific vectors remain directly comparable before aggregation, because they are produced by the same feature extractor and differ only in the spatial support used for pooling. This makes the subsequent scale-weighting module operate on a common representation basis rather than reconciling embeddings produced by independently trained branches.

3.3.2 PET/clinical-conditioned patient-specific scale fusion

Conventional attention in Eq. (6) determines aggregation weights from learnable scores. In RA-HMFNet, the score of each peritumoral scale additionally depends on the patient-level PET and clinical representations:

$$s_i = g_\theta\left([z_i; z_{\mathrm{PET}}; z_C]\right), \qquad i \in \{03, 36, 69\}. \tag{8}$$

The patient-specific scale weights and aggregated peritumoral representation are

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j\in\{03,36,69\}} \exp(s_j)}, \qquad z_{\mathrm{peri}} = \sum_{i\in\{03,36,69\}} \alpha_i z_i. \tag{9}$$

Thus, patients with similar CT scale representations may receive different peritumoral weights when their PET or clinical context differs. PET and clinical representations remain available for the subsequent multimodal fusion stage. The learned scale weights are treated as internal mixture coefficients rather than causal or biological importance measures. Conditioning is therefore asymmetric by design: PET and clinical context influence the weighting of peritumoral CT representations, but the scale-attention module does not replace their independent modality representations. This separation preserves modality-specific information for the second-stage fusion while allowing cross-modal context to modulate the within-CT aggregation step. Operationally, the scorer g_θ is applied to each scale-specific CT vector together with the same PET and clinical context. The resulting three scalar scores are normalized jointly, so increasing the relative contribution of one peritumoral scale necessarily reduces the relative contribution assigned to the others. The mechanism therefore changes the mixture over an already defined set of regional representations rather than generating a new spatial representation from scratch. Because the scale weights are non-negative and sum to one, the fused peritumoral representation is a convex combination of the three scale-specific representations. Appendix C further characterizes its deviation from fixed equal fusion and relates this deviation to the concentration of the learned scale weights.

3.3.3 Hierarchical multimodal fusion

The second stage integrates the intratumoral representation z_{IT}, the adaptively aggregated peritumoral representation z_{peri}, PET representation z_{PET}, and clinical representation z_C.

After modality-specific projections $\psi_m(\cdot)$,

$$\beta = \mathrm{softmax}\left(h([z_{\mathrm{IT}}; z_{\mathrm{peri}}; z_{\mathrm{PET}}; z_C])\right), \qquad z_{\mathrm{fuse}} = \sum_m \beta_m \psi_m(z_m). \tag{10}$$

A final linear classifier maps z_{fuse} to the recurrence logit. This hierarchy explicitly separates two representation problems: distance-specific aggregation within CT and multimodal integration across CT, PET, and clinical information. The first stage compresses multiple peritumoral scale vectors into one context-conditioned representation, whereas the second stage combines this summary with intratumoral, PET, and clinical embeddings. Compared with exposing all scale-specific vectors directly to the final fusion block, this imposes an explicit ordering on interactions: local spatial structure is resolved before full cross-modal integration. The resulting hierarchy also provides a natural correspondence between architecture and experiment: P09 tests whether explicit localization is sufficient, ROI Equal tests whether preserving three regional vectors without adaptive weighting is useful, ROI Global introduces learned but patient-invariant scale allocation, and the full model adds patient-conditioned scale allocation. The final multimodal stage is then evaluated on top of these progressively structured peritumoral representations. This factorization also constrains the representation presented to the second-stage multimodal fusion module: the number of peritumoral scales affects first-stage aggregation but not the dimensionality of the aggregated peritumoral input to the cross-modal stage. A formal comparison with direct joint fusion is provided in Appendix C.

4 Experiments

Experimental setup. The final RA-HMFNet uses BCEWithLogitsLoss with pos_weight=1. Thus, "unweighted" refers only to the classification class weight and does not imply equal P03/P36/P69 scale weights. No oversampling, WeightedRandomSampler, focal loss, or auxiliary attention/entropy objective is used; scale and modality attention are learned solely through the classification objective. Optimization uses AdamW with learning rate 10^{-4} and weight decay 10^{-4} (Loshchilov & Hutter, 2019). Training is capped at 60 epochs. Within each inner split, early stopping monitors validation BCE with patience 10, and the outer-final training duration is set to the median of the three inner best epochs. The 3D workload uses micro-batches of one with gradient accumulation, and Group Normalization replaces Batch Normalization throughout the final architecture.

Evaluation uses five repeats of five-fold stratified outer cross-validation. Within each outer-training partition, three-fold inner cross-validation is used for model selection and training-duration selection; the outer test fold is never used for tuning. Each patient receives one outer-test probability per repeat, and the primary patient-level prediction is the mean of the five probabilities. All main metrics are calculated from these 107 patient-level mean OOF predictions.

The main comparator set contains a clinical logistic regression model, a conventional Radiomics Fusion model integrating CT radiomics, PET radiomics/quantitative PET features, and clinical variables, a direct Multimodal Concat deep baseline, and a frozen Full Equal Multimodal baseline. Matched architectural controls are evaluated under the same final unweighted protocol. We report AUROC, AUPRC, Brier score, and binary log loss from patient-level mean OOF probabilities. Calibration is summarized by an assessment-only logistic recalibration intercept and slope, together with the mean predicted probability and observed event fraction. Because the recurrence prevalence is 25/107, AUPRC is reported alongside AUROC. All comparisons are interpreted descriptively as point estimates. The comparator set is deliberately heterogeneous. Clinical LR tests how much of the endpoint can be captured without imaging; Radiomics Fusion provides an engineered-feature reference; Multimodal Concat tests whether simply combining learned modality vectors is sufficient; and Full Equal Multimodal retains all modalities while removing adaptive scale weighting. Together with the matched ablations, these references distinguish gains associated with spatial representation from gains associated with fusion complexity. The matched sequence is therefore interpreted incrementally rather than as a set of unrelated models: NoROI to P09 tests explicit localization, P09 to ROI Equal tests distance decomposition,

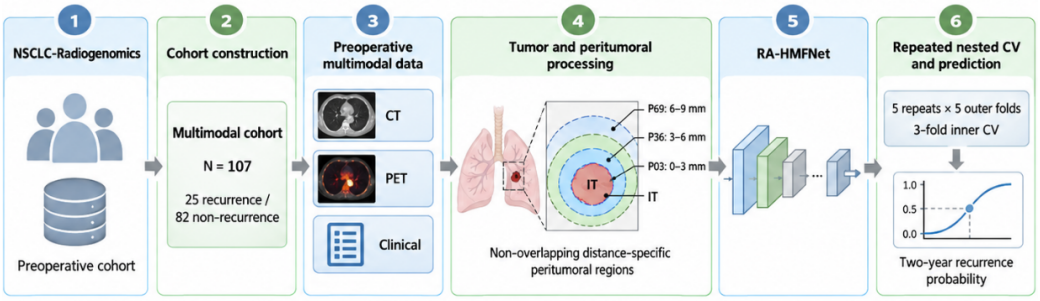


Figure 3: Study design and leakage-controlled evaluation framework. The final multimodal cohort contains 107 patients. All model selection is confined to training partitions, and final metrics use patient-level means of repeated outer-test OOF probabilities.

ROI Equal to ROI Global tests learnable scale weighting, and ROI Global to the full model tests patient-conditioned weighting. This ordering is used consistently when interpreting the numerical changes reported below. Given the limited cohort size (107 patients, 25 recurrence events), we avoid formal superiority claims and use repeated nested evaluation to reduce dependence on a single partition. The outer loop provides held-out estimates, while the inner loop is restricted to model-selection and training-duration decisions. This separation is especially important here because the same cohort is used for architecture development, ablation analysis, and final internal evaluation. For every outer split, all operations that require fitted parameters are confined to the corresponding training partition. This includes clinical preprocessing and all model-selection decisions. The outer test patients contribute only to held-out prediction for that split. Consequently, the five repeat-level predictions for a patient are generated by models for which that patient was excluded from the corresponding outer-training data. Repeating the outer folds yields multiple held-out predictions per patient; their mean is used as the single patient-level probability for the primary analysis. AUROC, AUPRC, Brier score, and log loss are therefore computed from the same 107 aggregated OOF probabilities, avoiding fold-wise metric averaging or selection of a favorable repeat.

4.1 Comparison with existing methods

Predictive performance. RA-HMFNet achieved an AUROC of 0.6634 and an AUPRC of 0.3861, with Brier score 0.1683 and log loss 0.5146. Among the principal baselines, Clinical LR achieved AUROC 0.6380, Radiomics Fusion 0.6434, Multimodal Concat 0.6000, and Full Equal Multimodal 0.6024. RA-HMFNet therefore had the highest AUROC and AUPRC point estimates in this comparator set and lower Brier and log-loss point estimates than these baselines. These comparisons are descriptive and are not interpreted as formal evidence of superiority.

Table 1: Main baseline comparison using patient-level mean repeated OOF predictions.

Model	Input / fusion	AUROC ↑	AUPRC ↑	Brier ↓	Log loss ↓
Clinical LR	Clinical	0.6380	0.3775	0.1758	0.5346
Radiomics Fusion	CT radiomics + PET radiomics/quantitative + clinical	0.6434	0.3409	0.1935	0.5812
Multimodal Concat	IT + Peri + PET + Clinical direct concatenation	0.6000	0.3258	0.2310	0.6561
Full Equal Multimodal	Full modalities; fixed equal P03/P36/P69 fusion	0.6024	0.3640	0.2365	0.6667
RA-HMFNet	ROI-aware adaptive hierarchical fusion	0.6634	0.3861	0.1683	0.5146

Probability quality and calibration. For the final RA-HMFNet, the Brier score was 0.1683 and log loss was 0.5146. The Brier skill score relative to the training-prevalence null benchmark was 0.0599. The assessment-only calibration intercept was −0.0347 and the slope was 0.9054. The mean predicted recurrence probability was 0.2212, compared with an observed recurrence fraction of 0.2336. These values do not indicate a large cohort-level probabil-

ity shift, but the sample size and absence of external validation do not support a claim of established calibration.

4.2 Matched architectural ablation

Variant definitions. NoROI removes ROI-aware peritumoral pooling and uses non-region-specific global aggregation. P09 uses a single ROI-aware 0–9 mm peritumoral representation without distance-specific decomposition. ROI Equal retains P03/P36/P69 pooling but fixes all scale weights to 1/3. ROI Global learns one three-element scale-weight vector shared across patients. RA-HMFNet uses PET/clinical-conditioned patient-specific scale weights. This matched sequence isolates spatial localization, distance decomposition, globally learnable weighting, and patient-specific weighting.

Component effects. Removing ROI-aware localization reduced the AUROC point estimate to 0.5810. Introducing a single ROI-aware P09 representation increased AUROC to 0.6478, an absolute numerical change of +0.0668 relative to NoROI. Decomposing P09 into P03/P36/P69 with fixed equal weights yielded AUROC 0.6390 and therefore did not improve the point estimate over P09. Learnable global scale weights increased AUROC from 0.6390 to 0.6566, and patient-specific adaptive weighting further increased it to 0.6634. The final numerical increment from ROI Global to RA-HMFNet was +0.0068, while the difference between RA-HMFNet and P09 was +0.0156.

Across the four probability metrics, RA-HMFNet had the highest AUROC and AUPRC, tied ROI Global for the lowest Brier score at four-decimal precision, while P09 had the lowest log loss (0.5127). The overall pattern therefore supports ROI-aware spatial localization as the largest numerical architectural effect in this cohort, but does not support the claim that distance-specific decomposition alone is sufficient to improve prediction. Importantly, the P09 control prevents the improvement of the full model from being attributed automatically to multiscale decomposition. Instead, the results suggest that localization provides the main representational benefit, whereas additional scale structure becomes useful only when accompanied by a learned aggregation rule.

Table 2: Final matched ablation results. All values are descriptive patient-level point estimates.

Model	ROI-aware	Distance-specific	Scale weighting	Patient-specific	AUROC ↑	AUPRC ↑	Brier ↓	Log loss ↓
NoROI	No	No	None	No	0.5810	0.3033	0.1777	0.5409
P09	Yes	No	Single 0–9 mm ROI	No	0.6478	0.3457	0.1689	0.5127
ROI Equal	Yes	Yes	Fixed equal	No	0.6390	0.3491	0.1697	0.5194
ROI Global	Yes	Yes	Learnable global	No	0.6566	0.3658	0.1683	0.5145
RA-HMFNet	Yes	Yes	Patient-specific adaptive	Yes	0.6634	0.3861	0.1683	0.5146

Attention behavior and optimization diagnostics. The learned peritumoral scale allocation was close to uniform. Mean normalized repeat-level entropy was 0.9923, and entropy of patient-mean weight vectors was 0.9964. The same top-weighted scale was selected in all five repeats for 40.19% of patients and in at least four repeats for 66.36%. Median pairwise repeat Spearman correlations were approximately 0.50–0.54, while ICC(3,1) ranged from 0.1718 to 0.2741 across the three scales. These findings indicate observable patient-level variation but limited repeat-level stability; the scale weights should therefore not be interpreted as stable biological biomarkers.

Optimization diagnostics also showed small-cohort overfitting pressure. Across 75 inner fits, the median best epoch was 9 and the median best validation BCE was 0.5123. The median final recorded train and validation BCE values were 0.3098 and 0.6604, respectively, with a median gap of 0.3476. Validation loss typically reached an early minimum and later rose or oscillated, supporting early stopping rather than additional capacity or longer training on this cohort.

5 Discussion

The matched experiments show that explicit peritumoral localization produced the largest numerical architectural effect. P09 exceeded NoROI by 0.0668 AUROC, whereas fixed equal decomposition into P03/P36/P69 did not improve over P09. The main empirical signal therefore comes from where deep features are pooled, not from distance decomposition alone.

Learnable weighting recovered higher point estimates than fixed equal fusion, with ROI Global reaching 0.6566 AUROC and the full patient-specific model 0.6634. However, the additional gain from patient-specific weighting was small (+0.0068), and the scale weights remained close to uniform. The attention mechanism should therefore be interpreted as a conditional aggregation component rather than evidence of stable patient-specific biological scale preference.

Clinical LR and Radiomics Fusion remained competitive in this cohort, while RA-HMFNet achieved the highest AUROC and AUPRC point estimates among the compared models and lower Brier and log-loss values than the principal deep baselines. These results support region-aware structured fusion as a useful inductive bias, but they do not establish statistical superiority or external generalizability.

From a representation perspective, the architecture separates three operations that are often conflated: spatial readout, within-CT scale aggregation, and cross-modal fusion. The matched controls make these stages experimentally distinguishable, and the analytical results in Appendix C clarify how merged-region pooling and adaptive convex aggregation constrain the resulting representations.

6 Limitations

The final cohort contains only 107 patients and 25 recurrence events, so performance point estimates may change materially in larger samples. All analyses use one public dataset, and no independent external cohort was available. The pipeline depends on tumor segmentation, PET SUV preprocessing, and PET-to-CT registration, which may introduce dataset-specific errors. Although RA-HMFNet computes patient-specific scale weights, these weights were near uniform and only moderately repeatable, so they should not be treated as biological explanations. Model comparisons are based on descriptive point estimates rather than formal inferential evidence of superiority. Finally, the fixed 730-day endpoint does not model recurrence timing beyond the horizon and should not be interpreted as a survival model.

Independent external validation of the frozen RA-HMFNet v1.0 configuration is therefore required before any claim of clinical applicability.

7 Conclusion

We introduced RA-HMFNet, a compact region-aware hierarchical multimodal framework for preoperative prediction of postoperative two-year recurrence in NSCLC. The method combines feature-map-level peritumoral ROI pooling, PET/clinical-conditioned scale aggregation, and hierarchical CT/PET/clinical fusion. In repeated nested internal validation, RA-HMFNet achieved AUROC 0.6634 and AUPRC 0.3861. Matched ablations showed that ROI-aware localization produced the largest numerical improvement, whereas fixed distance-specific decomposition alone did not improve over a single 0–9 mm peritumoral region. Learnable aggregation recovered higher discrimination, but patient-specific attention remained close to uniform and should not be interpreted causally. These findings motivate region-aware structured fusion while underscoring the need for independent validation before clinical use.

Reproducibility Statement

We report the cohort definition, preprocessing, model architecture, training protocol, model-selection procedure, and repeated nested cross-validation in the main paper. All reported metrics are computed from patient-level mean outer-test out-of-fold predictions.

AI Use Statement

Generative AI tools assisted with methodology discussion, code review, literature discovery, translation, and manuscript drafting. All experiments, numerical results, and final claims were executed or verified by the authors, who take responsibility for the final manuscript.

References

M. Amini, M. Nazari, I. Shiri, et al. Multi-level multi-modality (PET and CT) fusion radiomics: prognostic modeling for non-small cell lung carcinoma. Physics in Medicine & Biology, 66(20):205017, 2021. doi: 10.1088/1361-6560/ac287d.

Shaimaa Bakr, Olivier Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng, Jalen Anthony Benson, Weiruo Zhang, Ann N. C. Leung, et al. A radiogenomic dataset of non-small cell lung cancer. Scientific Data, 5:180202, 2018. doi: 10.1038/sdata.2018.202.

Tyler J. Bradshaw, Zachary Huemann, Junjie Hu, and Arman Rahmim. A guide to cross-validation for artificial intelligence in medical imaging. Radiology: Artificial Intelligence, 5(4):e220232, 2023. doi: 10.1148/ryai.220232.

Tugba Akinci D'Antonoli, Alessandra Farchione, Jacopo Lenkowicz, Marco Chiappetta, Giuseppe Cicchetti, Antonella Martino, Alessandra Ottavianelli, Riccardo Manfredi, Stefano Margaritora, Lorenzo Bonomo, Vincenzo Valentini, and Anna Rita Larici. CT radiomics signature of tumor and peritumoral lung parenchyma to predict nonsmall cell lung cancer postsurgical recurrence risk. Academic Radiology, 27(4):497–507, 2020. doi: 10.1016/j.acra.2019.05.019.

M. Kirienko, L. Cozzi, L. Antunovic, et al. Prediction of disease-free survival by the PET/CT radiomic signature in non-small cell lung cancer patients undergoing surgery. European Journal of Nuclear Medicine and Molecular Imaging, 45(2):207–217, 2018. doi: 10.1007/s00259-017-3837-7.

Soomin Lee, Julip Jung, Helen Hong, and Bong-Seog Kim. Prediction of two-year recurrence-free survival in operable NSCLC patients using radiomic features from intra- and size-variant peri-tumoral regions on chest CT images. Diagnostics, 12(6):1313, 2022. doi: 10.3390/diagnostics12061313.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2019.

G. Mehri-Kakavand, S. Mdletshe, M. Amini, and A. Wang. Multimodal radiomics fusion for predicting postoperative recurrence in NSCLC patients. Journal of Cancer Research and Clinical Oncology, 151(10):261, 2025. doi: 10.1007/s00432-025-06311-w.

S. B. Park, K.-U. Kim, Y. W. Park, J. H. Hwang, and C. H. Lim. Application of 18F-fluorodeoxyglucose PET/CT radiomic features and machine learning to predict early recurrence of non-small cell lung cancer after curative-intent therapy. Nuclear Medicine Communications, 44(2):161–168, 2023. doi: 10.1097/MNM.0000000000001646.

Yuxin Wu and Kaiming He. Group normalization. In European Conference on Computer Vision, pp. 3–19, 2018.

A Detailed Implementation Settings

The final model is implemented in PyTorch and outputs raw classification logits. Training applies BCEWithLogitsLoss directly to these logits with pos_weight=1, without an additional sigmoid before the loss. Three-dimensional training uses a micro-batch size of one with gradient accumulation over four micro-steps. AdamW is used without a learning-rate scheduler. Scale-attention and modality-fusion parameters are optimized only through the binary classification objective; no auxiliary attention, entropy, contrastive, or modality-consistency loss is used.

For each outer-training partition, the number of final training epochs is determined from the median of the three inner-fold best epochs. All model-selection and early-stopping decisions are restricted to the corresponding training partitions; the outer-test partition is used only for final prediction.

B Cohort and Preprocessing Details

The binary endpoint is defined relative to the surgery date. Patients with documented recurrence within 730 days after surgery are labeled positive. Recurrence occurring after 730 days is not counted as a positive event for this fixed-horizon endpoint. Cases without recurrence require sufficient follow-up to evaluate the 730-day endpoint; cases with insufficient follow-up, unavailable recurrence status, or inconsistent temporal information are excluded from endpoint construction.

The final multimodal cohort additionally requires a usable CT tumor segmentation, PET data passing the quantitative preprocessing and CT-registration pipeline, and availability of the selected preoperative clinical variables. The resulting cohort contains 107 patients, including 25 positive and 82 negative cases.

The seven tumor-location indicators used in the clinical branch correspond to right upper lobe, right middle lobe, right lower lobe, left upper lobe, left lower lobe, lingular location, and unknown location. Together with age at histological diagnosis, weight, sex, ethnicity, smoking status, pack-years, and quit-smoking year, these form the 14 preoperative clinical variables used by the model.

PET-to-CT alignment uses rigid six-degree-of-freedom registration with a mutual-information criterion before PET resampling to the CT geometry. Clinical preprocessing parameters that require fitting are estimated only from the corresponding training partition and then applied to validation or outer-test patients.

C Analytical View of Representation and Fusion

This section provides an analytical view of the representation and fusion operations used in RA-HMFNet. The goal is not to establish a generalization guarantee, but to make explicit the structural constraints induced by region-aware pooling, adaptive scale aggregation, and hierarchical multimodal fusion.

C.1 Region-aware representation decomposition

Let $F(v) \in \mathbb{R}^d$ denote the feature vector at spatial location v. Consider K mutually disjoint regions $\{R_i\}_{i=1}^K$ whose union is R, and assume that all pooled representations are computed from the same feature map and with the same averaging operator. Define

$$z_i = \frac{1}{|R_i|} \sum_{v \in R_i} F(v), \tag{11}$$

and let pooling over the union be

$$z_R = \frac{1}{|R|} \sum_{v \in R} F(v). \tag{12}$$

Because the regions form a non-overlapping partition of R,

$$z_R = \sum_{i=1}^{K} \frac{|R_i|}{|R|} z_i. \tag{13}$$

Thus, pooled representation over a merged region is a convex combination of the corresponding region-specific representations, with coefficients determined by regional support rather than by the prediction task.

For the three peritumoral regions used in RA-HMFNet, the idealized same-feature-map decomposition is

$$z_{\mathrm{P09}} = \pi_{03} z_{03} + \pi_{36} z_{36} + \pi_{69} z_{69}, \qquad \pi_i = \frac{|R_i|}{|R_{03}| + |R_{36}| + |R_{69}|}. \tag{14}$$

The significance of distance-specific pooling is therefore not that decomposition must create additional predictive information. Rather, it keeps $\{z_{03}, z_{36}, z_{69}\}$ explicitly available until a task-dependent aggregation rule is applied. Once these regions are pooled jointly, their individual representations are no longer separately accessible to subsequent fusion modules. This distinction motivates the matched P09 and distance-specific controls in the main experiments.

C.2 Properties of adaptive scale fusion

The scale-fusion stage has the form

$$z_{\mathrm{peri}}(x) = \sum_{i=1}^{K} \alpha_i(x) z_i, \qquad \alpha_i(x) \geq 0, \qquad \sum_{i=1}^{K} \alpha_i(x) = 1, \tag{15}$$

with $K = 3$ in the present implementation. Hence,

$$z_{\mathrm{peri}}(x) \in \mathrm{conv}\{z_1, \ldots, z_K\}, \tag{16}$$

so adaptive scale fusion performs a data-dependent interpolation among the available scale-specific representations rather than an unconstrained linear combination.

Fixed equal fusion is obtained with $u_i = 1/K$,

$$z_{\mathrm{eq}} = \sum_{i=1}^{K} u_i z_i, \tag{17}$$

whereas globally learned scale fusion corresponds to $\alpha_i(x) = \bar{\alpha}_i$ independent of the patient. Equal, global, and patient-specific fusion can therefore be viewed as nested forms of the same aggregation family with progressively more flexible weighting rules.

The representation difference between patient-specific and equal fusion is

$$z_{\mathrm{peri}}(x) - z_{\mathrm{eq}} = \sum_{i=1}^{K} \left(\alpha_i(x) - \frac{1}{K} \right) z_i. \tag{18}$$

If the scale-specific representations satisfy $\|z_i\| \leq B$ for all i, the triangle inequality gives

$$\|z_{\mathrm{peri}}(x) - z_{\mathrm{eq}}\| \leq B \, \|\alpha(x) - u\|_1 , \tag{19}$$

where $u = (1/K, \ldots, 1/K)$.

Let normalized attention entropy be

$$H_{\mathrm{norm}}(\alpha) = -\frac{1}{\log K} \sum_{i=1}^{K} \alpha_i \log \alpha_i. \tag{20}$$

Since

$$D_{\mathrm{KL}}(\alpha \| u) = (1 - H_{\mathrm{norm}}(\alpha)) \log K, \tag{21}$$

Pinsker's inequality yields

$$\|\alpha - u\|_1 \leq \sqrt{2(1 - H_{\text{norm}}(\alpha)) \log K}. \tag{22}$$

Combining the two bounds gives

$$\|z_{\text{peri}}(x) - z_{\text{eq}}\| \leq B\sqrt{2(1 - H_{\text{norm}}(\alpha(x))) \log K}. \tag{23}$$

Equation (23) gives a representation-level interpretation of the attention entropy reported in the main text. As normalized entropy approaches one, the maximum deviation from equal fusion contracts under the bounded-feature assumption. This does not imply equivalent predictive performance, because downstream nonlinear mappings may amplify small representation changes, but it limits the change introduced at the scale-fusion operation itself.

The same formulation also clarifies the difference between globally shared and patient-specific weighting. Let

$$z_{\text{global}}(x) = \sum_i \bar{\alpha}_i z_i. \tag{24}$$

Then

$$z_{\text{peri}}(x) - z_{\text{global}}(x) = \sum_i (\alpha_i(x) - \bar{\alpha}_i)\, z_i, \tag{25}$$

and therefore, under the same bounded-feature assumption,

$$\|z_{\text{peri}}(x) - z_{\text{global}}(x)\| \leq B\, \|\alpha(x) - \bar{\alpha}\|_1 . \tag{26}$$

This shows directly that patient-specific fusion can differ substantially from global weighting only when its learned scale allocation departs substantially from the shared global allocation or when the scale-specific representations themselves are large in norm.

C.3 Hierarchical fusion as structured factorization

Consider K peritumoral scale representations $z_1, \ldots, z_K \in \mathbb{R}^{d_p}$ together with intratumoral, PET, and clinical representations. A direct joint fusion mapping can be written abstractly as

$$z_{\text{direct}} = f_{\text{joint}}\left(z_{\text{IT}}, z_1, \ldots, z_K, z_{\text{PET}}, z_C\right). \tag{27}$$

If the non-peritumoral inputs have a combined dimensionality d_o, the dimensionality directly exposed to this joint cross-modal fusion function is

$$K d_p + d_o. \tag{28}$$

RA-HMFNet instead first constructs

$$z_{\text{peri}} = f_{\text{scale}}\left(z_1, \ldots, z_K \mid z_{\text{PET}}, z_C\right) \in \mathbb{R}^{d_p}, \tag{29}$$

and subsequently performs

$$z_{\text{fuse}} = f_{\text{modal}}\left(z_{\text{IT}}, z_{\text{peri}}, z_{\text{PET}}, z_C\right). \tag{30}$$

The complete mapping can therefore be expressed as the structured composition

$$f_{\text{RA}} = f_{\text{modal}} \circ f_{\text{scale}}, \tag{31}$$

with PET and clinical context participating in the first-stage weighting while remaining available to the second-stage multimodal fusion.

After scale aggregation, the peritumoral contribution to the second-stage cross-modal input is d_p, independent of K, rather than Kd_p. Relative to directly exposing all scale-specific representations to the multimodal stage, this reduces the dimensionality presented to the second-stage fusion function by

$$(K - 1)d_p. \tag{32}$$

This is not a claim that the complete RA-HMFNet necessarily contains fewer trainable parameters, because the first-stage scale-scoring module introduces its own parameters. Instead, the hierarchy constrains the interaction structure: relationships among distance-specific peritumoral representations are resolved before the resulting representation participates in full multimodal integration. It therefore acts as an architectural inductive bias about the order in which spatial and cross-modal information interact, rather than as a formal guarantee of improved generalization.