PHORA: Physiology-Guided Hierarchical Ordinal Radiomics for LI-RADS Classification in Multiphase CT

Dharini Raghavan[1], Advait Madabhushi[2], Amritpal Singh[2], Mendel Lebowitz[2], Gourav Modanwal[2], and Anant Madabhushi[2]

[1] Georgia Institute of Technology, Atlanta, GA, USA
draghavan7@gatech.edu (corresponding author)
[2] Emory University, Atlanta, GA, USA

Abstract. Automated Liver Imaging Reporting and Data System (LI-RADS) assessment from multiphase CT is not naturally a standard seven-class classification problem. LR-1–LR-5 form an ordered spectrum of increasing hepatocellular-carcinoma likelihood, whereas LR-M and LR-TIV represent clinically distinct outcomes that leave this ordinal continuum. We present PHORA, a physiology-guided hierarchical radiomics framework that models these two aspects of the task separately while retaining clinically recognizable intermediate evidence. PHORA combines conventional radiomics with spatial and temporal enhancement descriptors, distills training-only radiologist annotations into inference-time estimates of major LI-RADS concepts, and integrates special-category recognition with ordinal severity estimation through a clinically constrained router. On the 59-case public validation set, PHORA achieved a composite score of 0.718, adjusted quadratic weighted kappa of 0.738, and special-category recognition of 0.605. Among reference ordinal cases retained on the ordinal pathway, 30 of 31 predictions were within one LI-RADS category of the reference, indicating that residual staging errors were usually local rather than severe. The concept models recovered clinically important features such as non-rim arterial phase hyperenhancement and delayed washout with AUROCs of 0.915 and 0.905, respectively, while prediction accuracy decreased from 0.800 in the lowest-uncertainty quartile to 0.467 in the highest. These results show that PHORA provides a clinically structured and auditable formulation of LI-RADS prediction in which ordinal severity, special-category evidence, latent imaging concepts, routing logic, and model uncertainty can all be examined rather than collapsing the task into a single nominal class decision.

Keywords: Hepatocellular carcinoma · LI-RADS · Radiomics · Ordinal learning · Concept bottleneck · Multiphase CT

1 Introduction

Hepatocellular carcinoma (HCC) is unusual among solid tumors in that, for appropriately selected at-risk patients, characteristic findings on multiphase CT

or MRI can establish the diagnosis without tissue sampling. LI-RADS provides a standardized language for this assessment, assigning observations to LR-1 through LR-5 as the probability of HCC increases, while reserving LR-M for malignant observations that are not HCC-specific and LR-TIV for tumor in vein [2]. This structure has made LI-RADS attractive for computational decision support, and recent work has shown that automated analysis of major CT features and multiphase images can reproduce useful parts of the reporting process [8, 9].

The full seven-class problem remains difficult for two reasons. First, the labels do not have a single geometry. LR-1–LR-5 are ordered, so an LR-4/LR-5 disagreement is qualitatively different from an LR-1/LR-5 disagreement; ordinal-learning methods are designed precisely for settings in which the ordering of classes should influence the loss or decision rule [3]. LR-M and LR-TIV, however, are not simply more severe points on this scale. The AMPLIFAI challenge captures this distinction by combining adjusted QWK over the ordinal categories with balanced recognition of ordinal, LR-M, and LR-TIV cases [6]. A flat seven-way model can still perform well, but it does not make this separation explicit.

Second, LI-RADS reasoning is inherently multiphase. Arterial enhancement, washout, capsule appearance, lesion size, and targetoid or vascular patterns are interpreted together and often depend on where enhancement occurs within the lesion and how it evolves over time. Radiomics provides a natural way to quantify such heterogeneity from routine imaging and to retain information about texture, shape, and enhancement beyond visual assessment alone [7, 4]. At the same time, the AMPLIFAI training set contains radiologist-defined major-feature labels and masks that are not available on the private test set. They are therefore useful supervision, but they cannot simply be passed to the final classifier at inference.

PHORA (Physiology-guided Hierarchical Ordinal Radiomics Architecture) combines these ideas in one inference-safe pipeline: conventional radiomics and spatial/temporal physiology descriptors form the imaging representation, training-only annotations are distilled into predicted clinical concepts and burdens, and the final decision separates special-category routing from ordinal LR-1–LR-5 staging. To separate representation quality from decision architecture, the flat baselines use the same candidate imaging pool and training-only screening procedure; the multi-view stack provides a late-fusion alternative over the same feature families. We therefore evaluate PHORA with the official challenge score as well as ablations, concept recovery, class-specific behavior, ordinal error distance, uncertainty, repeatability, and source-shift analyses.

2 Materials and Challenge Task

2.1 AMPLIFAI data

The AMPLIFAI challenge comprises 764 multiphase CT studies: 531 training, 59 public validation, and 174 private test cases [6]. The publicly released training and validation data comprise 590 cases curated and harmonized from four contributing datasets. Each case includes an arterial phase and at least one later contrast phase,

with portal venous, delayed, and non-contrast acquisitions available according to the source protocol. A voxel-level target-lesion mask is supplied for every scored lesion. Training and validation metadata also contain LI-RADS major-feature labels and, when available, voxel masks for APHE, washout, and capsule.

Seventy training rows were labeled "No lesion", which is not one of the seven legal challenge outputs. The classification experiments therefore used the 461 training cases labeled LR-1–LR-5, LR-M, or LR-TIV; all 59 public-validation cases belonged to the same seven classes. Training counts were LR-1: 2, LR-2: 2, LR-3: 13, LR-4: 41, LR-5: 231, LR-M: 115, and LR-TIV: 57. Validation counts were 1, 1, 3, 6, 27, 14, and 7, respectively. Arterial CT was available for all 520 analyzed lesion cases, portal venous for 516, delayed for 388, and non-contrast for 365. Missing phases were represented explicitly rather than imputed by synthetic images.

2.2 Evaluation objective

Let $Y \in \{\mathrm{LR1}, \ldots, \mathrm{LR5}, \mathrm{LRM}, \mathrm{LRTIV}\}$. The official AMPLIFAI score is

$$S = 0.85\,\mathrm{QWK}_{\mathrm{adj}} + 0.15\,\mathrm{SCR}, \tag{1}$$

where adjusted QWK measures agreement on the LR-1–LR-5 scale and SCR is three-class balanced accuracy after collapsing predictions to {ordinal, LR-M, LR-TIV}. The metric gives most of its weight to ordered agreement while still requiring the two special categories to be recognized, which closely matches the hierarchical formulation used by PHORA.

3 Method

3.1 Overview and hierarchical formulation

Figure 1 summarizes the pipeline. Given a lesion mask Ω and phase image $I_p(v)$ for $p \in \{\mathrm{DRY}, \mathrm{ART}, \mathrm{VEN}, \mathrm{DEL}\}$, PHORA first extracts an inference-safe imaging representation $\phi(X, \Omega)$. Training-only annotations supervise auxiliary models for clinically recognizable concepts $\hat{z}$ and spatial annotation burdens $\hat{b}$; the ground-truth concept labels and masks are never supplied directly to validation or test cases.

The final classifier is organized as two linked decisions. A special-category branch estimates whether the lesion should remain on the ordinal LR-1–LR-5 continuum or be routed to LR-M or LR-TIV. Lesions retained on the ordinal branch receive a continuous severity estimate that is converted to LR-1–LR-5 with ordered cut-points. A final router reconciles these two sources of evidence and applies a small number of LI-RADS-motivated consistency rules.

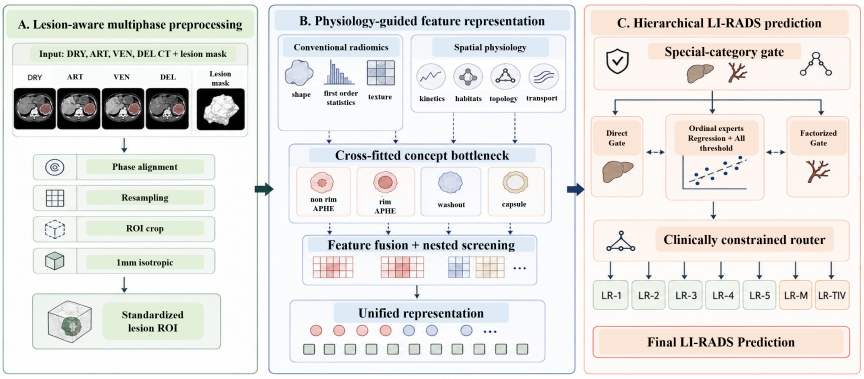

Fig. 1. Overview of PHORA, from multiphase CT and lesion mask to physiology-guided features, cross-fitted LI-RADS concepts, and hierarchical clinically constrained prediction.

3.2 Lesion-aware preprocessing and radiomics

PHORA uses two complementary image-processing views. The spatial-physiology view maps all available phases and the lesion mask to the arterial grid and retains physical spacing for geometric measurements. The conventional radiomics view performs phase-specific mask alignment, a 5-mm padded crop, and 1-mm isotropic resampling. CT interpolation is linear/B-spline as appropriate to the view, while masks use nearest-neighbor interpolation. Intensities are discretized with a fixed 25-HU bin width over $[-1024, 3071]$ HU. We compute 3-D shape and first-order statistics together with GLCM, GLRLM, GLSZM, GLDM, and NGTDM descriptors over 13 three-dimensional directions. These families follow standard radiomics practice [7, 4] and IBSI terminology [10]; our extractor is IBSI-inspired but has not been formally benchmarked against IBSI reference phantoms. Feature extraction succeeded for all 520 analyzed lesions.

3.3 Spatial and temporal physiology descriptors

A distance-to-boundary field divides each lesion into rim, middle, and core habitats, while a 6-neighbor lesion graph $G = (V, E)$ captures local spatial variation through

$$\mathrm{TV}(I) = \frac{1}{|E|} \sum_{(u,v)\in E} |I(u) - I(v)|, \qquad \mathcal{E}(I) = \frac{1}{|E|} \sum_{(u,v)\in E} \big(I(u) - I(v)\big)^2. \quad (2)$$

Directional semivariograms, Moran's I, Geary's C, and hotspot descriptors summarize larger-scale spatial organization. Temporal change is represented directly in Hounsfield units through

$$\begin{aligned} \Delta_{\mathrm{ART-DRY}} &= I_{\mathrm{ART}} - I_{\mathrm{DRY}}, & \Delta_{\mathrm{VEN-ART}} &= I_{\mathrm{VEN}} - I_{\mathrm{ART}}, \\ \Delta_{\mathrm{DEL-ART}} &= I_{\mathrm{DEL}} - I_{\mathrm{ART}}, & \Delta_{\mathrm{DEL-VEN}} &= I_{\mathrm{DEL}} - I_{\mathrm{VEN}}. \end{aligned} \quad (3)$$

Each field contributes whole-lesion and habitat statistics, graph energies, semivariograms, and fractions of voxels showing positive or negative change. Enhancement trajectories add dynamic range, path length, phase-index AUC, peak phase, end-to-start change, and curvature. A 3–12 mm perilesional reference ring, 1-Wasserstein/Jensen–Shannon phase distances, and compact vascular-contact and targetoid surrogates complete the approximately 900-variable candidate representation.

3.4 Cross-fitted concept and burden distillation

The training metadata define six binary auxiliary concepts: non-rim APHE, rim APHE, venous washout, delayed washout, venous capsule, and delayed capsule. For each concept j, an XGBoost classifier [1] estimates

$$\hat{z}_j = P(B_j = 1 \mid \phi). \tag{4}$$

This resembles a soft concept bottleneck [5]: the predicted concepts are exposed to the final model, but the original imaging features are retained rather than forcing every decision through the concepts alone. When a voxel-level feature mask M_j is available, its lesion-normalized burden is $b_j = |M_j \cap \Omega|/|\Omega|$, and a regression head estimates $\hat{b}_j$.

Leakage is avoided by assigning training cases only inner out-of-fold concept and burden predictions. Validation and test cases receive predictions from auxiliary heads fitted on the corresponding training partition. Predicted concepts, burden estimates, and lesion size are then combined into soft interaction features that capture clinically plausible relationships among APHE, washout, capsule, and size without ever using the true validation/test annotations.

3.5 Nested screening and hierarchical prediction

The candidate representation is large relative to the cohort, so screening is repeated within each training fold. Variables with at least 40% missing values or negligible variance are removed. The remainder are ranked using normalized univariate scores for three targets: the ordinal/LR-M/LR-TIV gate, ordinal severity, and available clinical concepts. High-ranking variables are greedily pruned for pairwise correlation at $|r| < 0.985$. The full configuration retains at most 140 imaging variables before the cross-fitted concept, burden, and interaction features are appended.

The coarse branch decision is represented in two complementary ways. A direct gate estimates $g_{\mathrm{dir}}(x)$ over {ordinal, LR-M, LR-TIV}. A factorized gate first estimates $p_S = P(S = 1 \mid x)$ for either special category and then $p_T = P(Y = \text{LR-TIV} \mid S = 1, x)$, yielding

$$P(\text{ordinal}) = 1 - p_S, \quad P(\text{LR-M}) = p_S(1 - p_T), \quad P(\text{LR-TIV}) = p_S p_T. \tag{5}$$

For ordinal lesions, PHORA combines a class-balanced XGBoost regressor with a shared all-threshold model estimating $q_k(x) = P(Y_o > k \mid x)$ for $k = 1, \ldots, 4$,

following the general principle of decomposing ordinal classification into ordered binary decisions [3]. The router combines branch and ordinal evidence as

$$g(x) = \beta g_{\text{dir}}(x) + (1-\beta) g_{\text{fac}}(x), \qquad s(x) = \alpha r(x) + (1-\alpha) a(x). \tag{6}$$

Four ordered cut-points map $s(x)$ to LR-1–LR-5. LR-M or LR-TIV is selected only when its probability clears both a category threshold and the ordinal probability by a margin. Predicted rim APHE smoothly reweights LR-M evidence. On the ordinal branch, LR-5 predictions below 10 mm or with sufficiently low predicted non-rim APHE are mapped to LR-4.

For the fixed public-validation experiment, $\beta = 0.25$ and $\alpha = 0.50$; the ordinal cut-points were 1.373, 3.009, 3.693, and 4.183; LR-M and LR-TIV thresholds were 0.327 and 0.699; the special-versus-ordinal margin was 0.080; the non-rim APHE threshold used by the LR-5 consistency rule was 0.291; and the rim-APHE LR-M reweighting coefficient was 0.971. These values were selected by randomized search over training out-of-fold predictions using the exact challenge score. No routing parameter was tuned on public validation.

4 Experiments

4.1 Protocol and comparison models

Architecture and feature-family ablations were evaluated with three-fold stratified cross-validation on the 461-case training subset. Because LR-1 and LR-2 were extremely rare, stratification labels were coarsened when necessary to preserve stable folds and representation of the special categories. After the method specification was fixed, models were trained on all 461 training cases and evaluated once on the 59-case public validation split. The final experiment used up to 140 screened imaging variables, ensemble seeds $\{17, 37, 61\}$, 5000 router-search trials, and 10,000 stratified bootstrap replicates.

The comparison models were chosen to separate the representation from the decision architecture. Elastic-net multinomial logistic regression, RBF-SVM, random forest, Extra Trees, histogram gradient boosting, and flat seven-class XGBoost all use the same inference-safe candidate feature pool and training-only screening procedure as PHORA. They therefore answer a controlled question: how well can the same quantitative representation perform when the final task is treated as ordinary nominal classification? A simple hierarchy uses a direct three-way gate followed by ordinal regression, but omits concept/burden distillation, the factorized gate, the all-threshold expert, and the clinical constraints. The cross-fitted multi-view stack trains separate XGBoost models on morphology, radiomics, custom physiology, and context/transport views, then combines their out-of-fold probabilities with a logistic meta-learner. Finally, a non-deployable oracle uses radiologist-provided major-feature metadata on validation and is included only to estimate the headroom associated with perfect concept availability.

Table 1. Public validation performance ($n = 59$). Flat classifiers and the late-fusion reference operate on the same inference-safe imaging representation used by PHORA; they differ primarily in how the final LI-RADS decision is formed. Oracle concepts use radiologist metadata at validation time and are therefore non-deployable.

Method	Composite	QWK	SCR
Flat nominal decision models			
Majority class	0.050	0.000	0.333
RBF-SVM	0.281	0.232	0.555
HistGradientBoosting	0.200	0.113	0.692
Elastic-net logistic	0.690	0.702	0.620
Flat XGBoost	0.687	0.686	0.692
Random Forest	0.705	0.709	0.683
Extra Trees	**0.763**	**0.785**	0.635
Structured and fusion references			
Simple hierarchical XGBoost	0.692	0.703	0.629
Cross-fitted multi-view stack	0.734	0.741	**0.694**
Proposed clinically structured formulation			
PHORA	0.718	0.738	0.605
Non-deployable upper bound			
Oracle concepts[†]	0.935	0.947	0.870

4.2 Secondary analyses

Architecture ablations remove concept and burden distillation, the factorized special gate, the all-threshold ordinal expert, the clinical constraints, nested feature screening, or burden distillation alone. Feature-family experiments evaluate morphology, handcrafted radiomics, custom kinetics/spatial descriptors, context/optimal-transport descriptors, all features except radiomics, and the full representation. Secondary analyses measure concept AUROC/AUPRC, burden correlation and MAE, per-class recall, gate/ordinal-expert uncertainty, repeatability across three CV seeds, train-to-validation shift using RBF-MMD and a two-sample classifier, and exploratory leave-one-source-out robustness. Uncertainty is treated as an audit signal rather than a calibrated abstention guarantee.

5 Results

5.1 Public validation performance

Table 1 shows that no single baseline dominates every part of the task. PHORA reached 0.718 composite with QWK 0.738 and SCR 0.605. Extra Trees had the highest deployable composite point estimate (0.763) and QWK (0.785), while the multi-view stack had the highest deployable SCR (0.694). HistGradientBoosting illustrates the tension particularly clearly: it reached SCR 0.692 and 0.644 accuracy, the same accuracy as Extra Trees, but its QWK was only 0.113. Flat XGBoost also achieved SCR 0.692 while falling below PHORA on QWK. These

differences matter because they correspond to different failure modes: some models recognize the broad special/ordinal branch relatively well while losing the ordered structure within LR-1–LR-5, whereas others preserve ordinal agreement but are less effective on special categories.

Extra Trees is also informative about the representation: it is trained on the same candidate imaging variables and screening procedure rather than on a separate image-processing pipeline. Its performance therefore shows that the PHORA feature space is strongly predictive with a generic ensemble. PHORA uses the same underlying information in a more explicit decision process that exposes special-category routing, ordinal severity, predicted concepts, and clinical consistency logic.

The 10,000-replicate stratified bootstrap gave a mean PHORA composite of 0.715 with 95% CI $[0.552, 0.853]$, QWK 0.734 $[0.541, 0.897]$, and SCR 0.605 $[0.469, 0.751]$. The paired PHORA-minus-Extra-Trees differences were -0.052 $[-0.129, 0.008]$ for composite, -0.056 $[-0.138, 0.008]$ for QWK, and -0.031 $[-0.182, 0.119]$ for SCR. The point estimate favors Extra Trees, but the width of these intervals reflects the uncertainty of ranking models on only 59 validation cases.

5.2 Class-specific and ordinal behavior

The public split is highly imbalanced. PHORA recall was 0/1 for LR-1, 0/1 for LR-2, 1/3 for LR-3, 1/6 for LR-4, 22/27 for LR-5, 10/14 for LR-M, and 2/7 for LR-TIV. The first four categories and LR-TIV therefore have too little support for stable class-wise conclusions. Once an ordinal case was routed correctly, however, the staging errors were usually small: among 31 reference LR-1–LR-5 cases retained on the ordinal pathway, 24 were exact, six differed by one level, and one differed by two or more levels. Thus 30/31 (96.8%) of retained ordinal predictions were within one LI-RADS category of the reference. This conditional analysis describes staging behavior and deliberately excludes the seven reference ordinal cases routed to a special category.

5.3 Architecture and feature analysis

The ablations show that the components of PHORA play different roles. Removing concept/burden distillation, nested screening, spatial-burden distillation, or the clinical constraints produced substantial decreases in the CV composite. In contrast, the scores obtained without the factorized gate or without the all-threshold expert were close to the full configuration. We therefore avoid interpreting every branch as an independently proven accuracy gain. The factorized and all-threshold views remain useful because they make special-category membership and ordered severity explicit, while the present cohort is too small to establish a separate gain from each branch.

The feature-family experiments lead to a similar conclusion. Handcrafted radiomics was the strongest isolated family at 0.536 composite, whereas the

Table 2. Architecture ablation on training CV. Each component was removed independently, so small differences around the full-model score should be interpreted cautiously.

Configuration	Composite	QWK	SCR
PHORA full	0.505	0.496	0.559
without concept + burden distillation	0.392	0.376	0.479
without factorized special gate	0.507	0.503	0.526
without all-threshold ordinal expert	0.509	0.497	0.572
without clinical constraints	0.336	0.293	0.581
without nested feature screening	0.403	0.380	0.537
without spatial-burden distillation	0.409	0.387	0.535

custom kinetics/spatial family was weaker when used by itself. The full candidate representation also did not outperform radiomics in a naive early-fusion comparison. In PHORA, the spatial-physiology descriptors are therefore used as complementary information for concept prediction, local context, and structured interactions rather than as a replacement for conventional radiomics. This result is consistent with the broader radiomics literature, where feature reproducibility, dimensionality control, and careful validation are often as important as the size of the raw feature bank [10].

5.4 Clinical concepts, uncertainty, and robustness

Several of the auxiliary concept heads recovered clinically recognizable LI-RADS features on validation. AUROC/AUPRC were 0.915/0.960 for non-rim APHE, 0.702/0.245 for rim APHE, 0.870/0.787 for venous washout, 0.905/0.835 for delayed washout, 0.854/0.731 for venous capsule, and 0.718/0.164 for delayed capsule. Predicted burden also tracked the available feature masks: Spearman ρ was 0.404 for APHE, 0.625 for venous washout, 0.331 for delayed washout, and 0.432 for venous capsule, with corresponding MAEs of 0.229, 0.188, 0.234, and 0.082. The non-deployable concept oracle reached 0.935 composite, suggesting that improved estimation of the same clinically meaningful intermediate evidence could provide substantial headroom without changing the basic task formulation.

PHORA's internal uncertainty was associated with reliability. Accuracy was 0.800 in the lowest-uncertainty quartile and 0.467 in the highest, with intermediate quartiles of 0.533 and 0.643. The relationship is not perfectly monotone, which is unsurprising with roughly 15 cases per quartile, but the extremes separate a relatively reliable group from a substantially more difficult group. This makes uncertainty useful as an audit signal and a possible basis for selective human review, while stopping short of claiming a calibrated clinical reject option.

Across CV seeds 42, 142, and 242, composite/QWK/SCR were 0.453 ± 0.048, 0.434 ± 0.056, and 0.558 ± 0.001. Selected-feature diagnostics did not indicate a strong global train-to-validation shift (RBF-MMD^2 = 0.0071; two-sample AUC 0.461). Exploratory source-holdout scores were more variable (0.236, 0.678, and 0.327 across the three large source batches), indicating that acquisition

heterogeneity remains a relevant limitation even when the global shift statistic is small.

6 Discussion

PHORA was motivated by the structure of LI-RADS rather than by architectural complexity. An observation can move along the LR-1–LR-5 continuum, or it can leave that continuum because the imaging favors LR-M or LR-TIV. The benchmark shows why this distinction matters: methods with similar accuracy or SCR can have very different QWK. Extra Trees gives the strongest public-validation point estimate, but because it uses the same inference-safe candidate imaging pool and screening procedure, it also confirms that the PHORA representation itself is strongly predictive. PHORA uses the same information in a more explicit way, keeping the ordinal/special route, predicted APHE/washout/capsule evidence, and separate branch and severity scores visible. The paired bootstrap interval crosses zero, so the 59-case public split is too small for a strong population-level ranking between these decision rules.

The secondary analyses help explain what the structured formulation adds. When an ordinal lesion is routed correctly, 30/31 predictions are within one category of the reference; several concept heads recover major LI-RADS features with useful discrimination; and the lowest-uncertainty quartile is substantially more accurate than the highest. At the same time, the ablations argue for restraint: concept/burden distillation, screening, burden features, and clinical constraints show clear drops when removed, whereas the factorized gate and all-threshold expert do not show an independent score advantage in this small CV experiment. Conventional radiomics remains the strongest isolated feature family, with the custom physiology descriptors serving mainly as complementary information for concepts, context, and structured interactions.

The main limitation is sample size. LR-1 and LR-2 each contain only two training cases and one validation case, and LR-3, LR-4, and LR-TIV are also sparse, so rare-class recall is descriptive. The radiomics implementation is IBSI-inspired but not formally benchmarked against reference phantoms; concept heads inherit annotation uncertainty; and some vascular/targetoid descriptors are imaging surrogates rather than explicit anatomy. Routing thresholds were learned within the AMPLIFAI training distribution and would require recalibration elsewhere. PHORA is therefore best viewed as a structured decision framework built on a representation whose predictive value is independently supported by strong flat and late-fusion references, with larger and more balanced cohorts needed for prospective validation.

7 Conclusion

PHORA combines an inference-safe multiphase CT representation with cross-fitted LI-RADS concepts and a hierarchy that separates special-category routing from LR-1–LR-5 severity. Extra Trees achieved the highest public-validation

point estimate on the same candidate representation, while PHORA retained clinically recognizable intermediate outputs, local ordinal errors, and uncertainty that identified more difficult cases. Larger cohorts and external calibration are needed to establish rare-class and prospective clinical performance.

External-resource disclosure. No external patient datasets, private institutional data, pretrained or foundation models, pseudo-labels, synthetic images, or external annotations were used. All imaging and annotation supervision came from the AMPLIFAI public release; no external network access is required at inference.

Disclosure of interests. The authors declare no competing interests relevant to this work.

Acknowledgments. We thank the AMPLIFAI organizers and annotating radiologists for the challenge data and evaluation framework.

References

1. 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)
2. Chernyak, V., Fowler, K.J., Kamaya, A., et al.: Liver Imaging Reporting and Data System (LI-RADS) version 2018: Imaging of hepatocellular carcinoma in at-risk patients. Radiology 289(3), 816–830 (2018)
3. Frank, E., Hall, M.A.: A simple approach to ordinal classification. In: De Raedt, L., Flach, P. (eds.) ECML 2001. LNCS, vol. 2167, pp. 145–156. Springer, Berlin, Heidelberg (2001)
4. Gillies, R.J., Kinahan, P.E., Hricak, H.: Radiomics: Images are more than pictures, they are data. Radiology 278(2), 563–577 (2016)
5. Koh, P.W., Nguyen, T., Tang, Y.S., et al.: Concept bottleneck models. In: Proceedings of the 37th International Conference on Machine Learning. PMLR 119, 5338–5348 (2020)
6. Kulkarni, P., Shah, N., Suryavanshi, A., et al.: AMPLIFAI: A multiphase CT dataset for benchmarking clinical reasoning in LI-RADS assessment of liver lesions. arXiv:2608.14778 (2026)
7. Lambin, P., Rios-Velazquez, E., Leijenaar, R., et al.: Radiomics: Extracting more information from medical images using advanced feature analysis. European Journal of Cancer 48(4), 441–446 (2012)
8. Mulé, S., Ronot, M., Ghosn, M., et al.: Automated CT LI-RADS v2018 scoring of liver observations using machine learning: A multivendor, multicentre retrospective study. JHEP Reports 5(10), 100857 (2023)
9. Xu, Y., Zhou, C., He, X., et al.: Deep learning-assisted LI-RADS grading and distinguishing hepatocellular carcinoma from non-HCC based on multiphase CT: A two-center study. European Radiology 33(12), 8879–8888 (2023)
10. Zwanenburg, A., Vallières, M., Abdalah, M.A., et al.: The Image Biomarker Standardization Initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 295(2), 328–338 (2020)