This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Clinically-grounded Multiphasic Modeling for Interpretable LI-RADS Categorization

Muhammad Alberb[1,3], Helen M. C. Cheung[2,3], and Anne L. Martel[1,3,4]

[1] Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada
[2] Department of Medical Imaging, University of Toronto, Toronto, ON, Canada
[3] Sunnybrook Research Institute, Toronto, ON, Canada
[4] Vector Institute, Toronto, Ontario, Canada
muhammad.alberb@mail.utoronto.ca

Abstract. Hepatocellular carcinoma can be diagnosed directly from multiphasic CT via the Liver Imaging Reporting and Data System (LI-RADS). However, evaluating subtle phase-dependent enhancement patterns causes substantial inter-reader variability. Conventional deep learning models typically regress LI-RADS categories directly, discarding the guidelines' rule-based structure and proving highly vulnerable to the severe class imbalances inherent in clinical cohorts. To overcome this, we propose a transparent pipeline that explicitly decouples imaging feature prediction from deterministic clinical decoding. At its core, an efficient, context-aware multiphasic vision transformer confines attention to clinically relevant anatomy. A domain-grounded routing mechanism then pools these spatiotemporal embeddings into specific tissue regions to predict individual LI-RADS features. Finally, a probabilistic decoding strategy marginalizes these predictions over the official LI-RADS table. By embedding explicit clinical knowledge into the architecture, our method ensures interpretability and robustness. Our experiments revealed that our proposed multiphasic encoder equipped with the clinically-grounded decoder outperforms alternative methods by over 12 points in overall composite score. Ultimately, our approach achieved a composite score of 0.60 on the external test set, securing first place on the official AMPLIFAI challenge leaderboard at the time of submission. To support reproducibility, our code and models are publicly available[5].

Keywords: LI-RADS · Hepatocellular Carcinoma · Multiphasic CT

1 Introduction

Hepatocellular carcinoma (HCC) is the most common primary liver cancer and the third leading cause of cancer-related mortality worldwide [14]. Early detection of HCC can dramatically change outcomes, potentially increasing 5-year survival rate from under 20% to over 70% [13]. One unique characteristic of HCC is that it can often be definitively diagnosed in at-risk populations without tissue biopsy, relying entirely on specific contrast-enhancement patterns observed

[5] https://github.com/martellab-sri/auto-lirads

in multiphasic computed tomography (CT) [12]. Because the imaging assessment itself dictates the diagnosis and subsequent therapy, it demands exceptional interpretative consistency. To standardize this critical process and guide clinical management, the Liver Imaging Reporting and Data System (LI-RADS) was established to classify hepatic observations into distinct categories, ranging from definitely benign (LR-1) to definitely HCC (LR-5), alongside labels for probable or definite malignancy not specific to HCC (LR-M) and tumor in vein (LR-TIV) [3]. However, the high volume of longitudinal surveillance imaging required for at-risk patients places a substantial cognitive workload on radiologists, underscoring the clinical imperative for automation to maintain diagnostic consistency at scale.

Beyond alleviating this clinical workload, automation directly addresses the inherent subjectivity of visual assessment. Although LI-RADS provides a deterministic rule-based framework for diagnosis, extracting the requisite imaging features is highly complex. Radiologists routinely process multiple unaligned contrast phases and evaluate subtle relative intensity changes. This involves distinguishing malignant predictors, such as non-rim arterial phase hyperenhancement (APHE), subsequent washout appearance, and the presence of an enhancing capsule, as well as benign indicators like peripheral nodular enhancement (PNE). These phase-dependent judgments are further confounded by the characteristics of the liver parenchyma, which fluctuates with the timing of acquisition and the physiology of the patient. Consequently, inter-reader variability remains a recognized limitation, with reported poor-to-moderate LI-RADS categorization reliability [5]. By standardizing the extraction of these complex spatiotemporal features, automated models can eliminate the primary source of human variance, fully realizing the objective of a strictly reproducible imaging-based diagnostic standard.

In medical image classification, conventional deep learning approaches typically regress categorical labels directly from images without intermediate extraction of predefined features [10]. When applied to LI-RADS, this end-to-end paradigm discards the rule-based structure of the clinical guidelines [4,16]. Requiring a model to implicitly relearn this established diagnostic algorithm from data is inefficient and highly susceptible to the small sample size and severe class imbalance inherent in clinical cohorts [8]. Shifting the learning objective to the underlying imaging features leverages denser and more balanced supervision, guarantees adherence to clinical guidelines, and allows for interpretable categorization. Consequently, in this work, we approach automated LI-RADS categorization through independent feature prediction followed by deterministic decoding. To predict these features, we develop an anatomy-aware dynamic spatiotemporal encoder for dense multiphasic feature extraction, followed by a clinically-grounded task-aware decoder for APHE, PNE, washout, and enhancing capsule detection. Finally, LI-RADS scores are generated from these predicted features following the clinical algorithm. Our primary contributions are:

– A transparent pipeline that integrates deep representations with domain knowledge to ensure accurate and interpretable LI-RADS categorization.

- An efficient context-aware multiphasic encoder that restricts feature extraction to clinically-relevant regions.
- Extensive experiments demonstrating the benefits of integrating explicit domain knowledge into the modeling pipeline.
- Empirical validation through the AMPLIFAI challenge [9], where our method achieved first place on the official leaderboard at the time of submission.

2 Materials and Methods

2.1 Datasets

We developed and evaluated our approach using two multiphasic cohorts: the AMPLIFAI challenge dataset [9,7] and the MCT-LTDiag dataset [15].

AMPLIFAI. The dataset provides 520 cases, each containing a single annotated target lesion, an arterial phase, and at least one late phase. Reference standards include voxel-level segmentation masks for the lesion and its major LI-RADS features. Reflecting a high-risk population, the diagnostic categories are heavily skewed toward definite disease, comprising 258 LR-5, 129 LR-M, 64 LR-TIV, 47 LR-4, and 16 LR-3 cases. Conversely, benign representations are exceedingly rare, with only 3 LR-1 and 3 LR-2 cases. Major feature annotations include non-rim APHE in 291 cases, rim APHE in 69 cases, washout in 291 cases, and enhancing capsule in 134 cases.

MCT-LTDiag. To compensate for the sparse benign representation in the primary cohort, we utilized an auxiliary dataset of 517 studies distributed across four lesion types: 218 metastases, 103 hepatocellular carcinomas, 100 intrahepatic cholangiocarcinomas, and 96 haemangiomas. Alongside these subtypes, the dataset provides annotations for PNE, which is present in 80 of the benign cases. We leverage this cohort to pretrain our encoders and train a dedicated PNE detection head, thereby explicitly supplying the network with positive evidence of benign entities.

2.2 Pipeline

Our method formulates LI-RADS categorization as explicit feature prediction followed by fixed rule-based decoding rather than end-to-end direct classification. As show in Fig. 1, the pipeline comprises four stages: (i) context-aware multiphasic encoding, (ii) lesion subtype classification, (iii) clinically-grounded lesion-level detection of primary LI-RADS features, and (iv) rule-based marginalized LI-RADS decoding. To prevent overfitting on the highly imbalanced target dataset, the encoder is frozen after pretraining on subtype classification.

Context-Aware Multiphasic Encoding. We employ a phase-agnostic vision transformer backbone pretrained with self-supervised learning (SSL) using 3DINO [17]. To inject anatomical context, we incorporate a dedicated segmentation module into our encoder [1]. Predicted liver segmentation is then used to constrain self-attention to liver patches. This approach anchors feature extraction to the clinically-relevant anatomy while enabling efficient processing of

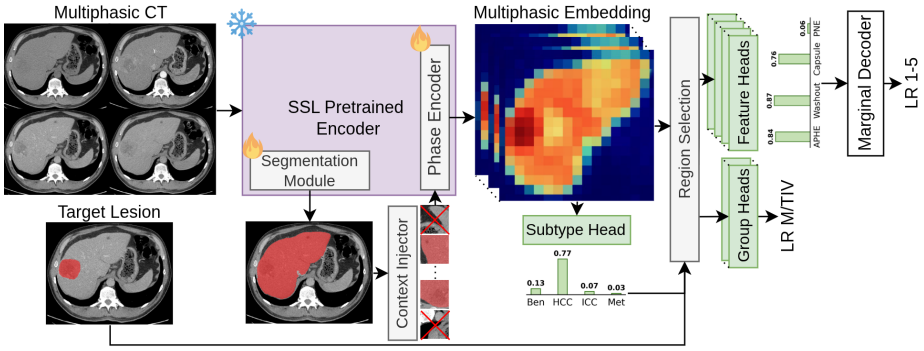

Fig. 1. Method Overview. A context-aware multiphasic encoder extracts spatial features, which are forwarded to clinically-grounded task-aware feature heads. The predicted features are then marginalized over the LI-RADS table to generate the final LI-RADS category.

multiple phases by drastically reducing sequence length within the backbone. Because the pretrained backbone is phase-agnostic, each available contrast phase is embedded independently. To capture contrast enhancement patterns, a patch-wise phase encoder with dynamic positional embeddings is incorporated to capture enhancement dynamics from available phases, inherently accommodating missing-phase scenarios.

Subtype Classification. Prior to LI-RADS categorization, the backbone and temporal encoder undergo subtype pretraining on the auxiliary cohort to distinguish HCC lesions from benign (Ben), intrahepatic cholangiocarcinoma (ICC), and metastatic (Met) lesions. The predicted subtype probabilities are then used as priors in the subsequent LI-RADS categorization stages.

Feature Detection. From the spatial embeddings generated by the multiphasic encoder, we utilize lesion segmentations to extract region representations via selective patch masking and aggregation. While our segmentation module supports automated lesion delineation, the dataset evaluates only a single pre-selected target lesion per scan. To prevent target-mismatch errors and align strictly with the challenge protocol, which provides manual masks for these targets during testing, we utilize the ground-truth segmentations directly at inference. Embeddings are pooled for four regions: whole lesion, lesion core after 5 mm erosion, lesion rim after subtracting the core, and liver parenchyma after subtracting 10 mm dilated lesions. Each feature detection head is then provided with specific region representations corresponding to clinical reading practices: APHE receives arterial core and rim representations together with their difference against the pre-contrast phase. Similarly, PNE receives core and rim but differencing against rim only. Enhancing capsule receives venous and delayed rim together with their difference against arterial rim. Washout receives venous and delayed core together with their difference against arterial core. When both venous and delayed phases are available, evidence from both phases is aggregated. For the group

categories, the LR-M head receives whole-lesion representations from arterial, portal venous and delayed phases together with venous-arterial difference, while the LR-TIV head receives the whole-lesion representation of the available late phases. Finally, all heads are injected with predefined intensity and morphology features computed directly from the masks and native Hounsfield units (HU), including shape, size, volumetric fractions, and per-phase mean attenuations, as well as prior subtype probabilities.

LI-RADS Decoding. Predicted features are mapped to a final category using the LI-RADS diagnostic table. Rather than applying hard decision thresholds, the table is marginalized over the predicted feature probabilities. Each configuration of APHE, washout, and enhancing capsule contributes to its corresponding cell, weighted by its probability and the measured lesion size. Because the standard table only resolves the ordinal categories LR-3 to LR-5, category assignment proceeds in two stages. First, the total ordinal probability mass is compared against the LR-M and LR-TIV probabilities to determine the diagnostic branch, and the most probable ordinal cell is selected only if the ordinal branch prevails. To correct for potential miscalibration of the LR-M and LR-TIV heads, two scale factors, τ_M and τ_{TIV}, rebalance the distribution, subject to a lower bound on ordinal recall. Finally, because the table encodes the hallmarks of HCC rather than positive evidence of a benign entity, categories LR-1 and LR-2 are not directly reachable. To address this, any case routed to the ordinal branch is explicitly reassigned to LR-1 when PNE is detected, or to LR-2 if washout is simultaneously detected.

2.3 Experimental Setup

Evaluation. Performance is evaluated using the official challenge composite score, defined as $0.85 \times \text{QWK} + 0.15 \times \text{SCR}$, where QWK denotes the adjusted quadratic-weighted kappa computed exclusively on cases that are both annotated and predicted as ordinal categories, and SCR denotes special category recognition, which measures the balanced accuracy across three diagnostic branches: ordinal, LR-M, and LR-TIV. Additionally, the predictive power and calibration of individual feature heads is assessed using the area under the receiver operating characteristic curve (AUC), balanced accuracy (Acc), and Brier score. Internal performance is estimated over the entire cohort using out-of-fold predictions. This ensures that each case is scored exclusively by models and thresholds optimized without access to its data. To distinguish genuine performance differences from partition noise, the evaluation procedure is repeated across 10 random seeds. We report the mean and standard deviation across these repetitions and evaluate configurations using paired comparisons on matched seeds. Finally, external validity is assessed on the held-out challenge test set, which is kept private from participants and scored independently via the official leaderboard.

Baselines and Ablations. We compare our approach against three families of baselines, all evaluated under the same protocol and on identical cross-validation partitions, where applicable. To evaluate our representation strategy, we deploy the original 3DINO backbone without subtype adaptation, temporal encoding,

or anatomical masking. To isolate the contribution of our learned features, a second baseline is restricted entirely to the predefined scalars, establishing the standalone predictive capacity of native attenuation, size, and shape descriptors. Finally, to assess our decoding method, we replace the fixed diagnostic table with a flat seven-way classifier and three two-stage alternatives. In these two-stage models, the ordinal branch is predicted using either a multinomial classifier, CORAL [2], or CORN [11], with the latter two specifically enforcing an explicit ordinal structure across categories LR-1 through LR-5. Beyond these structural baselines, we ablate our clinically motivated input routing by replacing the feature-specific region assignments with uniform designs where every predictive head receives identical data. Specifically, we evaluate providing all heads with the whole lesion across all phases, every region across every phase, explicit background features of liver parenchyma, and exclusively providing the arterial phase, which is the only phase present in all cases across the internal and external evaluation sets.

Implementation details. Volumes are resampled to 1mm isotropic resolution and cropped to a 112^3 window centred on the lesion, with intensities clipped to $[-175, 250]$ HU and rescaled to $[-1, 1]$. Low-rank adaptation (LoRA) [6] is applied to the pretrained backbone, with rank 32 for the multiphase encoder and rank 64 for the segmentation module. Each feature block is mean-imputed, standardised and reduced to 32 principal components, with a univariate filter retaining 128 components for the capsule head and 64 for each group head. The heads are L_2-regularised logistic regression models with inverse-frequency class weights, where the penalty is selected via three-fold inner cross-validation within a five-fold outer cross-validation repeated over ten seeds. The decoding gates are searched over $[0.20, 3.00]$ under an ordinal-recall bound of 0.8, and the benign override fires above a conservative predicted PNE probability threshold of 0.75.

3 Results

3.1 Feature Encoder

To evaluate the efficacy of our representation learning strategy, we compare our subtype-adapted encoder against the unadapted baseline (3DINO) and a representation-free baseline (*None*) that relies solely on native HU, size, and shape descriptors (Table 1). Relying exclusively on predefined scalar features yields poor overall LI-RADS performance (Composite = 0.350), confirming that geometry and attenuation alone cannot capture the complex dynamic enhancement of hepatic lesions. While the unadapted 3DINO backbone improves overall performance (Composite = 0.448), our multiphasic, context-aware, and subtype-adapted encoder achieves substantial gains (Composite = 0.584). This overall performance gap stems from two distinct factors. First, a severe discrepancy emerges during cross-cohort transfer. When the PNE head, which is identically trained for both encoders on the auxiliary MCT-LTDiag cohort, is applied to the primary AMPLIFAI dataset, the unadapted representation fails to transfer its operating point. Consequently, the benign override falsely triggers on 35 cases,

compared to just a single case for our adapted encoder. Second, the advantage of our subtype-adapted encoder extends beyond mitigating these transfer failures. Even when the benign override is entirely removed and the comparison is restricted to the 514 non-benign cases, our multiphasic encoder still outperforms the unadapted 3DINO backbone (Composite = 0.535 compared to 0.477). This persistent gap highlights a key property of our deterministic decoding strategy: because final categorization requires integrating multiple predicted probabilities, even slight, seemingly marginal improvements across individual feature heads compound into substantial gains in the final LI-RADS composite score. Ultimately, context-awareness and subtype adaptation enhance both predictive accuracy and generalizability.

Table 1. Our context-aware multiphasic encoder outperforms encoder-free approaches using intensity and morphology features (None) as well as the phase- and task-agnostic baseline (3DINO). LI-RADS, APHE, washout, and capsule performance is scored on AMPLIFAI (n=520), while PNE is scored on a held out MCT-LTDiag subset (n=171)

Encoder	LI-RADS			APHE	Washout	Capsule	PNE
	Composite	QWK	SCR	AUC / Acc	AUC / Acc	AUC / Acc	AUC / Acc
None	.350 ± .04	.316	.539	.904 / .834	.823 / .752	.699 / .649	.950 / .887
3DINO	.448 ± .04	.414	**.642**	**.921** / .845	.921 / .837	.759 / .688	.974 / .928
Ours	**.584 ± .03**	**.580**	.607	.919 / **.847**	**.929** / **.845**	**.773** / **.696**	**.986** / **.946**

3.2 Decoding Strategy

We next evaluate our fixed probabilistic decoding against end-to-end learned decoders (Table 2). To ensure a fair comparison, all alternative classifiers receive the complete union of feature blocks used across our dedicated heads. However, all learnable decoders, including direct end-to-end classification using a flat seven-class classifier, 3-class SCR classifier followed by a 5-class ordinal multinomial classifier, and SCR classifier followed by a CORAL or CORN ordinal classifiers, achieve inferior performance compared to our decoding strategy, which marginalizes predicted features over the deterministic LI-RADS diagnostic table. This demonstrates that explicitly embedding clinical guidelines provides far stronger regularization than end-to-end label regression, especially on imbalanced cohorts. To further validate the decoder marginalization, calibration was assessed via Brier score, yielding 0.12, 0.10, 0.18, and 0.02 for APHE, washout, capsule, and PNE, respectively. These scores indicate adequate calibration to support marginalization. A qualitative example of decoding is shown in Fig. 2b.

3.3 Feature Head Inputs

To validate our domain-grounded input routing, we compare our task-specific region selection against uniform input designs where every predictive head receives

Table 2. Performance of alternative decoding strategies. Alternatives replace the diagnostic table with a learned decoder using the same multiphasic features.

Decoder	Composite	QWK	SCR
flat 7-class	.404 ± .054	.369	.606
3-class + 5-class	.364 ± .047	.323	.601
3-class + CORAL	.454 ± .056	.424	.620
3-class + CORN	.397 ± .034	.358	**.622**
Ours	**.584 ± .025**	**.580**	.607

identical inputs (Table 3). Models restricted to whole-lesion representations or arterial-phase data alone achieve limited overall performance (Composite $\leq$ 0.430). Incorporating all available contrast phases and anatomical regions slightly improves performance (Composite = 0.464), underscoring the necessity of multiphasic anatomy-aware features. Routing the union of our clinically-motivated regions to every head in a task-agnostic manner further improves performance (Composite = 0.517). However, tailoring input regions to match clinical reading guidelines for each feature in a task-specific manner yields the highest performance (Composite = 0.584). Again, this highlights the importance of injecting domain knowledge as a prior into automated LI-RADS categorization.

Table 3. Ablation study of feature heads inputs. Evaluated regions include the whole lesion (L), core (C), rim (R), and background parenchyma (P). We also compare using all available phases to arterial phase only (Art). Task-agnostic represents using the union of our clinically-motivated regions for all heads.

Phases	Regions	LI-RADS			APHE	Washout	Capsule	PNE
		Composite	QWK	SCR	AUC / Acc	AUC / Acc	AUC / Acc	AUC / Acc
Art	L	.411 ± .05	.386	.550	.916 / .849	.836 / .789	.741 / .668	.985 / .930
All	L	.401 ± .02	.370	.575	.913 / .839	.876 / .794	.747 / .679	.984 / .945
All	L, P	.411 ± .05	.385	.560	.899 / .831	.874 / .788	.754 / .692	.980 / .930
Art	C, R, L, P	.430 ± .04	.404	.578	.918 / .849	.841 / .791	.745 / .681	.986 / .933
All	C, R, L, P	.464 ± .03	.445	.567	.915 / .840	.893 / .798	.764 / **.703**	.983 / .940
Task-agnostic		.517 ± .05	.503	.597	**.927** / **.857**	.927 / .832	.772 / .696	.985 / **.952**
Ours		**.584 ± .03**	**.580**	**.607**	.919 / .847	**.929** / **.845**	**.773** / .696	**.986** / .946

3.4 Challenge Results

On the held-out AMPLIFAI challenge test set, our method ranked first on the official leaderboard at the time of submission, achieving a composite score of 0.60, a QWK of 0.65, and an SCR of 0.34. While the test QWK exceeds our cross-validated internal estimate (QWK = 0.58), demonstrating strong generalization,

the reported SCR is significantly lower. This asymmetry reflects a deliberate operating point selection, as illustrated in Fig. 2a. By sweeping the lower bound on ordinal recall during gate optimization, QWK increases monotonically from 0.559 to 0.595 as more cases are retained within the ordinal branch, whereas SCR decreases from 0.603 to 0.498. Rather than selecting an operating point that optimizes SCR, we prioritized ordinal recall to maximize the primary component of the challenge metric, which weights QWK at 85%. Because QWK is evaluated exclusively on cases predicted and annotated as ordinal, prioritizing ordinal recall evaluates QWK over a larger and more representative subset of cases. The reported test SCR of 0.34 directly reflects this operating trade-off, securing peak performance on the overall challenge objective.

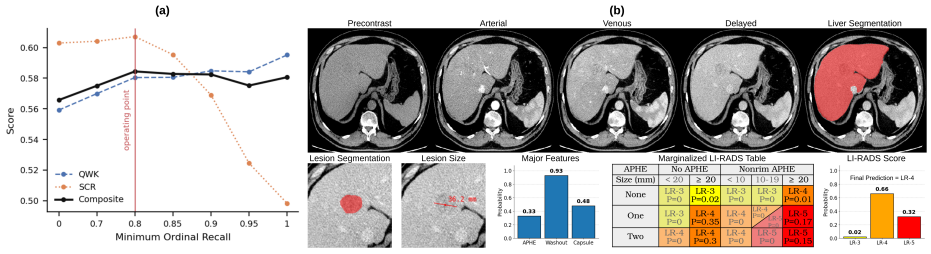

Fig. 2. (a) QWK-SCR tradeoff, showing the optimal operating point on internal data (red line). For the challenge, we instead prioritized maximizing ordinal recall to boost the QWK subset, which comprises 85% of the ranking metric. **(b)** Qualitative example of our probabilistic decoding, illustrating how predicted probabilities of major features are propagated through the marginal LI-RADS table to derive final LR-3, LR-4, and LR-5 probabilities. *None*, *One*, and *Two* indicate whether zero, either, or both additional features (washout, capsule) are present.

4 Conclusion

In this work, we presented a transparent, clinically-grounded pipeline for automated LI-RADS categorization of hepatic lesions from multiphasic CT. Recognizing the limitations of end-to-end deep learning on highly imbalanced clinical cohorts, we explicitly decoupled the diagnostic process into anatomy-aware multiphasic feature extraction and deterministic rule-based decoding. Our experiments demonstrated that integrating domain knowledge, through context-aware multiphasic encoding, subtype pretraining, task-specific anatomical input routing, and probabilistic marginalization over the LI-RADS diagnostic table, substantially outperforms alternative methods. By shifting the learning objective from direct category regression to interpretable clinical feature prediction, our approach enforces adherence to established radiological guidelines. This transparency is critical for clinical translation, as it empowers radiologists to verify

the specific anatomical evidence driving automated categorization rather than relying on a black-box prediction. Achieving first place on the AMPLIFAI challenge leaderboard at the time of submission further underscores the robustness of this framework to unseen data, demonstrating its potential to standardize HCC surveillance and reduce radiologist workload in clinical practice. Although this framework establishes a competitive benchmark, clinical deployment requires further optimization and validation across larger cohorts. Furthermore, although the underlying architecture inherently supports comprehensive surveillance via multi-lesion analysis, the current evaluation is constrained by the challenge protocol to a single pre-localized lesion. Finally, the extreme scarcity of benign instances in the primary cohort induces metric instability and necessitates an operational trade-off between ordinal recall and special category recognition. Addressing these data constraints remains a vital path forward to fully realize the clinical potential of guideline-adherent AI.

Acknowledgments. This work was supported by the Society of Abdominal Radiology (SAR) through the Wylie J. Dodds Research Award (2021), as well as the Canada Foundation for Innovation (Grant No. 40206), and the Ontario Research Fund. The research was also enabled in part by support provided by Compute Ontario (computeontario.ca) and the Digital Research Alliance of Canada (alliancecan.ca)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alberb, M., Chen, J., El-rewaidy, H., Karanicolas, P., Seth, A., Amemiya, Y., Martel, A., Cheung, H.: An automated radiomics framework for postoperative survival prediction in colorectal liver metastases using preoperative mri (2026), https://arxiv.org/abs/2603.10216
2. Cao, W., Mirjalili, V., Raschka, S.: Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognition Letters **140**, 325–331 (2020). https://doi.org/10.1016/j.patrec.2020.11.008
3. Chernyak, V., Fowler, K.J., Kamaya, A., Kielar, A.Z., Elsayes, K.M., Bashir, M.R., Kono, Y., Do, R.K., Mitchell, D.G., Singal, A.G., Tang, A., Sirlin, C.B.: Liver imaging reporting and data system (li-rads) version 2018: Imaging of hepatocellular carcinoma in at-risk patients. Radiology **289**(3), 816–830 (2018). https://doi.org/10.1148/radiol.2018181494, pMID: 30251931
4. Hamm, C.A., Wang, C.J., Savic, L.J., Ferrante, M., Schobert, I., Schlachter, T., Lin, M., Duncan, J.S., Weinreb, J.C., Chapiro, J., Letzen, B.: Deep learning for liver tumor diagnosis part i: development of a convolutional neural network classifier for multi-phasic mri. European Radiology **29**(7), 3338–3347 (Jul 2019). https://doi.org/10.1007/s00330-019-06205-9
5. Hong, C.W., et al.: A multicenter assessment of interreader reliability of li-rads version 2018 for mri and ct. Radiology **307**(5), e222855 (2023). https://doi.org/10.1148/radiol.222855, pMID: 37367445

6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
7. Kulkarni, P., Shah, N., Suryavanshi, A., Delfino, J., Tonascia, J., Wong-You-Cheong, J., Lane, B., Chirico, J., Hirsch, J.D., Li, A., Huang, H., Doo, F.X.: AMPLIFAI: A multiphase CT dataset for benchmarking clinical reasoning in LI-RADS assessment of liver lesions. Preprint (2026), https://um-ihc-ca2i.github.io/amplifai-challenge/assets/2026_KulkarniShah_AMPLIFAI_preprint.pdf
8. Piffer, S., Ubaldi, L., Tangaro, S., Retico, A., Talamonti, C.: Tackling the small data problem in medical image classification with artificial intelligence: a systematic review. Progress in Biomedical Engineering **6**(3), 032001 (jun 2024). https://doi.org/10.1088/2516-1091/ad525b
9. Shah, N., Kulkarni, P., Suryavanshi, A., Guo, J., Delfino, J., Lane, B., Wong-You-Cheong, J., Huang, H., Doo, F.: Amplifai: Annotated multi-phase liver imaging for artificial intelligence (Apr 2026). https://doi.org/10.5281/zenodo.19848293
10. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annual Review of Biomedical Engineering **19**(Volume 19, 2017), 221–248 (2017). https://doi.org/10.1146/annurev-bioeng-071516-044442
11. Shi, X., Cao, W., Raschka, S.: Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. Pattern Analysis and Applications **26**(3), 941–955 (Aug 2023). https://doi.org/10.1007/s10044-023-01181-9
12. Singal, A.G., Llovet, J.M., Yarchoan, M., Mehta, N., Heimbach, J.K., Dawson, L.A., Jou, J.H., Kulik, L.M., Agopian, V.G., Marrero, J.A., Mendiratta-Lala, M., Brown, D.B., Rilling, W.S., Goyal, L., Wei, A.C., Taddei, T.H.: Aasld practice guidance on prevention, diagnosis, and treatment of hepatocellular carcinoma. Hepatology **78**(6), 1922–1965 (May 2023). https://doi.org/10.1097/HEP.0000000000000466
13. Singal, A.G., Pillai, A., Tiro, J.: Early detection, curative treatment, and survival rates for hepatocellular carcinoma surveillance in patients with cirrhosis: A meta-analysis. PLOS Medicine **11**(4), 1–20 (04 2014). https://doi.org/10.1371/journal.pmed.1001624
14. Sung, H., Filho, A.M., Laversanne, M., Ferlay, J., Siegel, R.L., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2024: Globocan estimates of incidence and mortality worldwide for 34 cancers in 186 countries. CA: A Cancer Journal for Clinicians **76**(4), e70090 (2026). https://doi.org/10.3322/caac.70090
15. Wu, X., Su, H., Hua, Y., Xu, Y., Wang, L., Wang, X., Wang, S., Jin, B., Liu, X., Wan, X., Sun, Q., Wang, X., Du, S.: A multi-phase ct dataset for automated differential diagnosis of liver tumors. Scientific Data **13**(1), 31 (Dec 2025). https://doi.org/10.1038/s41597-025-06343-4
16. Wu, Y., White, G.M., Cornelius, T., Gowdar, I., Ansari, M.H., Supanich, M.P., Deng, J.: Deep learning li-rads grading system based on contrast enhanced multiphase mri for differentiation between lr-3 and lr-4/lr-5 liver tumors. Annals of Translational Medicine **8**(11) (2020). https://doi.org/10.21037/atm.2019.12.151
17. Xu, T., Hosseini, S., Anderson, C., Rinaldi, A., Krishnan, R.G., Martel, A.L., Goubran, M.: A generalizable 3d framework and model for self-supervised learning in medical imaging. npj Digital Medicine (2025). https://doi.org/10.1038/s41746-025-02035-w