Shortcut-Aware Validation for LI-RADS Categorisation on Multi-Phase CT

Harsh Vardhan, Sakshi Suryawanshi, and Subhamoy Mandal (✉)

School of Medical Science and Technology,
Indian Institute of Technology Kharagpur, Kharagpur, India
{harshvardhan25t, s.suryawanshi25}@kgpian.iitkgp.ac.in
smandal@smst.iitkgp.ac.in

Abstract. LI-RADS categorises focal liver lesions in at-risk patients on an ordinal scale from definitely benign to definitely hepatocellular carcinoma, and the assigned category informs clinical management. The AMPLIFAI challenge benchmarks automating this on multi-phase CT, scoring a composite of adjusted quadratic weighted kappa (QWK) and special-category recognition. Cross-validation scores on its public corpus far exceed private-test performance—a gap usually blamed on distribution shift. We show that two non-transferable shortcuts account for much of it. First, slice thickness identifies the source cohort, whose label mixes are near-disjoint—the 0.8 mm stratum is 52% special-category, the 2.5 mm stratum 79% LR-5—so acquisition alone recognises the special categories well above chance, but carries no ordinal signal. Second, 70 lesion-free examinations are scored as LR-1 and 69 ship no segmentation, so lesion diameter—zero exactly when the mask is absent—identifies the anchoring benign class: that scalar alone reaches QWK 0.888, and 0.871 composite with acquisition added—within 0.04 of a full 3D CNN on the same folds, without reading a voxel. Every private case supplies a mask, so none of this transfers. Our cohort-grouped, masked-only protocol addresses both and discriminates where the public one cannot: a baseline scoring QWK 0.943 publicly has no ordinal agreement at all, QWK 0.000. Under it, a method built from the LI-RADS definitions—peri-lesional relative normalisation, region-structured pooling, dense feature supervision and phase-difference channels—scores 0.41 on 174 private cases from three unseen centres, where error analysis localises the residual deficit to small-volume LR-TIV read as LR-M. Both diagnostics are inexpensive and worth running on any harmonised benchmark.

Keywords: LI-RADS · Multi-phase CT · Ordinal classification · Shortcut learning · Challenge validation

1 Introduction

Hepatocellular carcinoma (HCC) is one of the few solid tumours diagnosed from imaging alone [6], and the Liver Imaging Reporting and Data System (LI-RADS) formalises that pathway [2]: an at-risk patient's observation receives an ordinal

category from LR-1 (definitely benign) to LR-5 (definitely HCC), set by non-rim arterial phase hyperenhancement (APHE), non-peripheral washout, enhancing capsule and size, alongside the separate categories LR-M (probably malignant, not HCC-specific) and LR-TIV (tumour in vein). Because the category drives management, an automated system must reproduce an ordinal clinical decision.

AMPLIFAI [5] evaluates this on multi-phase contrast-enhanced CT: a model receives the phase volumes and a segmentation of the target lesion, and returns one category. Its public corpus of 590 cases is harmonised from four sources [1,3,7,10]; the held-out set is 174 cases from three centres of one previously unseen health system. Submissions score

$$S = 0.85 \cdot \kappa_{\text{adj}} + 0.15 \cdot \text{SCR}, \tag{1}$$

where κ_{adj} is quadratic weighted kappa over the ordinal categories, restricted to cases whose ground truth *and* prediction are ordinal, and SCR is three-class balanced accuracy over {ordinal, LR-M, LR-TIV}. The two components are near-orthogonal: LR-M and LR-TIV predictions are excluded from κ_{adj} entirely.

We start from an important observation: cross-validation scores on the public corpus far exceed anything achieved on the private set—ours by a wide margin, and no leaderboard entry exceeds $S = 0.48$ while public scores above 0.90 are easy to obtain (Sect. 2). The usual explanation is distribution shift. We propose a different one and test it.

Hypothesis. *If the public corpus contains features that predict the label but cannot transfer, then a model denied all image content should still score well under the standard protocol, and a protocol that removes those features should separate models the standard protocol cannot.*

Both halves are testable before any private evaluation, and Sect. 2 confirms them; such confounding is usually diagnosed only after deployment, but here it is diagnosable in advance.

Related work. Challenge rankings are known to be unstable under small changes to metric or aggregation [8]; we address a complementary failure, where the *data* admit near-perfect scores without solving the task. The mechanism is shortcut learning [4], in medical imaging often acquisition-related [12]—but diagnosed there with external data *after* training, whereas we identify it from released metadata beforehand, which matters because the default recipe of selecting variants by cross-validation selects on the shortcut here.

Contributions. *(i)* We identify two shortcuts in the public AMPLIFAI corpus and separate their effects: acquisition leaks the special categories, mask absence the ordinal scale (Sect. 2). *(ii)* We propose a cohort-grouped, masked-only protocol under which a conventional baseline is revealed to have zero ordinal agreement (Sect. 2.3). *(iii)* We develop a LI-RADS-motivated framework under it, evaluate it on the held-out set, and localise its residual failure mode (Sects. 3–6).

2 Two Shortcuts in the Public Benchmark

2.1 Acquisition Almost Determines the Label

The corpus merges four sources with differing acquisition protocols and non-comparable clinical populations. Source identity is not released but survives harmonisation in the reconstruction geometry: slice thickness is strongly discrete—317 cases at exactly 0.8 mm and 177 at exactly 2.5 mm—so binning at its empty gaps recovers four cohorts of 320, 47, 190 and 33 cases, each absorbing a few near-neighbouring values.

Those cohorts have almost disjoint clinical compositions. The 0.8 mm cohort ($n = 320$) is 15% LR-5 and 52% special-category, and alone holds 115 of the 129 LR-M, 53 of the 64 LR-TIV and 68 of the 73 benign observations; the 2.5 mm cohort ($n = 190$) is 79% LR-5 and 8% special-category. Thickness alone separates LR-5 from the rest with an area under the ROC curve of 0.82.

What this leaks, however, is the *special* categories rather than the ordinal ladder. A gradient-boosted classifier given only in-plane spacing and slice thickness reaches three-class balanced accuracy 0.510 over {ordinal, LR-M, LR-TIV}, well above the 0.333 of chance, but quadratic weighted kappa exactly 0.000: acquisition says which *cohort* a case came from, and so how likely LR-M or LR-TIV is, but carries no ordinal information. Its cost is paid in SCR and in making cross-cohort transfer unmeasurable (Sect. 6).

2.2 Mask Absence Identifies the Benign Class

The corpus contains 70 examinations in which no target observation was recorded. The official metric scores these as LR-1, alongside only three cases carrying an explicit LR-1 label, and 69 of the 70 have no lesion segmentation—not an oversight but a consequence of the label, since an examination with no observation has nothing to segment. The benign end of the scale is thus identifiable without reference to image content.

This is decisive for κ_{adj}, and it is measurable. The supplied lesion diameter is 0 for 68 of the 69 unmasked cases and at least 9.9 mm for all 521 masked ones, so diameter alone encodes mask absence: that scalar reaches $\kappa_{\text{adj}} = 0.888$, and 0.871 composite with the acquisition scalars added—within 0.04 of every image-based variant in Table 1. Essentially all ordinal agreement without reading a voxel comes from mask absence, not acquisition, because LR-1 anchors the scale and quadratic weighting punishes distant confusions most. Yet every test case ships a segmentation, so we evaluate on the 521 masked cases too—a diagnostic for deployment, not a replacement for the official task definition; excluding the rest costs 12% of the corpus and almost the entire benign class (Sect. 6).

2.3 A Protocol That Predicts

Our primary protocol combines both corrections: stratified group five-fold cross-validation grouped so an acquisition signature—the pair of in-plane spacing and slice thickness—never spans a split, evaluated only on masked cases, scored with

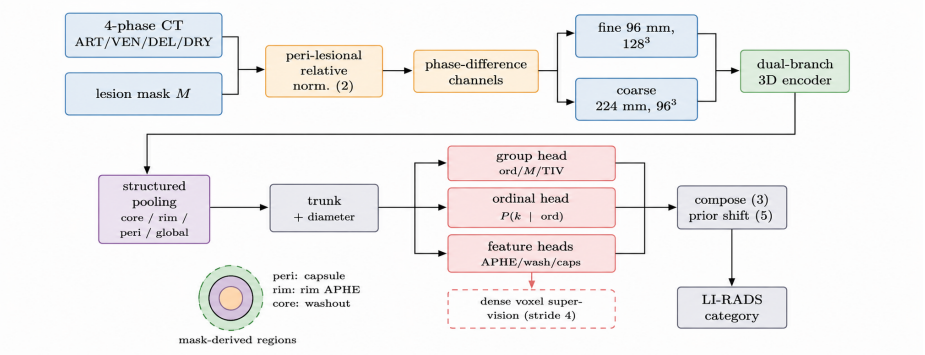

Fig. 1. Pipeline. Four phases and the lesion mask are normalised against peri-lesional parenchyma, augmented with phase-difference channels and cropped at two fixed physical extents. Two 3D encoder branches feed region-structured pooling; three heads are composed into a seven-class posterior and corrected for prior shift. Inset: the three mask-derived regions, each the location at which a LI-RADS feature is defined

the official implementation. The two corrections do different work. Grouping leaves the probe untouched (0.870 ungrouped, 0.871 grouped)—its power is diameter, not an acquisition parameter—and constrains only the acquisition leak, paid in SCR rather than κ_{adj}. The masked-only restriction is what removes the ordinal shortcut, cutting the probe to 0.459. A metadata model still reaches $\kappa_{\text{adj}} = 0.450$ there, but that residue is legitimate: diameter is itself a LI-RADS criterion (Sect. 4), whereas mask absence is an artefact, and it is the artefact the restriction removes. One signature (0.8×0.78 mm) covers 52% of the corpus and cannot be held out intact, so it is randomly partitioned into five subgroups: grouping is exact for the other 106 signatures, approximate for that one. Because that split is random rather than by an independent covariate, cohort-grouping only approximately constrains acquisition-related leakage within this dominant stratum, and should not be read as fully removing it there.

Its value is clearest on the baseline. A conventional dual-scale 3D CNN reaches $\kappa_{\text{adj}} = 0.943$ scored on all 590 cases and $\kappa_{\text{adj}} = 0.000$ on the masked subset—no ordinal agreement at all once the benign class must be recognised from appearance, and below the three-scalar probe—so the official scoring cannot separate it from a model with genuine skill, both exceeding 0.94.

3 Method

Every component below follows from one observation: *LI-RADS is defined by where enhancement occurs and what it is relative to, never by absolute appearance.* APHE means brighter than liver; washout means darker than liver later; rim versus non-rim is location, not magnitude. A network pooling globally over absolute intensities discards this, and the components below each restore one part of it (Fig. 1).

3.1 Preprocessing

Peri-lesional relative normalisation. Absolute Hounsfield values encode scanner calibration, contrast dose, injection timing and habitus, all differing between the public cohorts and an unseen institution. For each phase p we take μ_p as the median intensity of a peri-lesional shell—the mask dilated by six voxels minus the lesion, restricted to $[-20, 300]$ HU so vessels, fat and air do not contaminate it—and normalise

$$\tilde{x}_p = \mathrm{clip}\left(\frac{x_p - \mu_p}{\sigma},\, -4,\, 4\right), \qquad \sigma = 60 \text{ HU}. \tag{2}$$

Every voxel is now expressed relative to the patient's own liver, which is how a radiologist reads the study.

Fixed physical field of view. Two cubes centred on the lesion centroid are extracted at fixed extent, 96 mm at 128^3 for detail and 224 mm at 96^3 for context. A lesion-scaled crop would encode size as resolution and entangle it with appearance; a fixed extent keeps physical size visible, with diameter supplied separately as a scalar.

Phase-difference channels. Washout is by definition cross-phase, and since each phase is liver-relative after (2) the three differences $\tilde{x}_{\mathrm{ART}} - \tilde{x}_q$, for $q \in \{\mathrm{VEN}, \mathrm{DEL}, \mathrm{DRY}\}$, are directly interpretable and scanner-invariant. Each is gated to zero when its operand phase is missing. Only 56% of cases have all four phases, so a presence indicator accompanies the input and phases are dropped at random in training.

3.2 Anatomically-Structured Pooling

Non-rim and rim APHE differ not in whether the lesion enhances but *where*: non-rim enhancement fills the core, rim enhancement the outer shell. Measured over all supplied voxel annotations, washout is 92.6% intralesional ($n = 412$), whereas an enhancing capsule straddles the boundary, 61.8% inside and 38.2% outside ($n = 164$). A globally averaged feature vector destroys all of this.

We therefore pool encoder features over four regions derived from M by morphological erosion ε and dilation δ (3^3 and 5^3 kernels at feature-map resolution, Fig. 1 inset): $R_{\mathrm{core}} = \varepsilon(M)$, $R_{\mathrm{rim}} = M \setminus \varepsilon(M)$, $R_{\mathrm{peri}} = \delta(M) \setminus M$ and $R_{\mathrm{glob}} = \Omega$, concatenating the region-averaged descriptors $z = \left[\, \bar{f}_R \,\right]_R$. The rim-versus-core contrast is then available to the heads directly rather than reconstructed from a mean.

3.3 Hierarchical Heads and Training Objective

The two components of (1) are near-orthogonal, so we factorise the output to match: a group head predicts $P(g)$ over {ordinal, LR-M, LR-TIV}, and an ordinal head the conditional ladder $P(k \mid \mathrm{ord})$ over LR-1 to LR-5. The seven-class posterior is

$$\begin{aligned} P(\text{LR-}k) &= P(g{=}\mathrm{ord})\, P(k \mid \mathrm{ord}), \quad k = 1, \ldots, 5, \\ P(\text{LR-M}) &= P(g{=}\mathrm{M}), \qquad P(\text{LR-TIV}) = P(g{=}\mathrm{TIV}). \end{aligned} \tag{3}$$

This keeps the two metric components separable at inference.

A third group of heads predicts the LI-RADS imaging features. APHE is three-way {absent, rim, non-rim} rather than binary, and among the cases in which it was assessed the three-way split is close to decisive on its own: all 69 showing rim APHE are LR-M, all 258 LR-5 cases show non-rim APHE, and all 73 benign cases show none. Voxel annotations supply positives as a mask and assessed negatives as an all-zero target, and supervise a decoder at stride four; they are training supervision only, and no feature annotation is needed at inference. The case-level logit is the mean of the $k = 32$ highest-scoring voxels inside the *dilated* lesion—top-k rather than mean keeps focal features such as a thin capsule from being averaged away, and the dilation keeps a capsule lying outside the boundary in view. APHE is not assessed for the 64 LR-TIV cases, so their targets are masked out of the APHE and dense losses rather than treated as negatives—a gap with a measurable cost (Sect. 5). The objective is

$$\mathcal{L} = \mathcal{L}_{\text{grp}} + \mathcal{L}_{\text{ord}} + \tfrac{1}{2}\mathcal{L}_{\text{emd}} + \tfrac{3}{5}\mathcal{L}_{\text{aphe}} + \tfrac{3}{10}\big(\mathcal{L}_{\text{wash}} + \mathcal{L}_{\text{caps}} + \mathcal{L}_{\text{case}}\big) + \mathcal{L}_{\text{dense}}, \quad (4)$$

with class-balanced cross-entropies (label smoothing 0.05) on the group and ordinal heads, an Earth-mover term enforcing ordinality, cross-entropy on APHE, binary cross-entropy on washout and capsule, a voxel-level dense loss, and $\mathcal{L}_{\text{case}}$ keeping the top-k readout calibrated against the case-level feature labels. Every feature term is masked to the cases in which that feature was assessed.

Augmentation against the acquisition shortcut. Because acquisition parameters predict the label, augmentation targets them directly. Effective through-plane resolution is randomised by average-pooling and repeating along a random axis (p=0.3, factor 2–4), alongside intensity scaling (±15%), additive Gaussian noise ($\sigma \leq 0.1$), axis flips (p=0.5 each), 90° axial rotations, and phase dropout (p=0.15 per phase).

3.4 Inference and Decision Rule

Predictions are averaged over four axis flips and five seeded models. Training is dominated by LR-5 (44%), whereas the private cohort is described as consecutive clinical examinations spanning LR-1 to LR-5. Since κ_{adj} is sensitive to the label marginal, we apply a prior-shift correction

$$P'(y) \;\propto\; P(y)\left(\frac{\pi_{\text{target}}(y)}{\pi_{\text{train}}(y)}\right)^{\alpha}, \quad (5)$$

with $\alpha = 0$ recovering the uncorrected model. Because π_{target} is unknown we evaluate against five plausible scenarios—the empirical masked-only prior (needing no shift, $\alpha = 0$) plus four alternative targets (surveillance-like, flat-ordinal, HCC-heavy, few-special), each realised by importance-resampling the out-of-fold cases. The grid is $\alpha \in \{0, 0.25, 0.5, 0.75, 1.0\}$ combined with each alternative target and two group biases, each on a 9-point grid over $[-1, 1]$; we take the parameters minimising the worst-case shortfall across all five scenarios against the best attainable in any one. Verified nested by fold, this lifts variant C from 0.530 at the argmax (Table 1) to 0.555.

4 Experiments

4.1 Setup

All models are trained on the 590 public cases under the protocol of Sect. 2.3, every metric computed by the official implementation. The encoder has two independent branches, one per crop scale, each a stem and four residual stages (channels 24–192, strides 1/2 to 1/32); 8.65 M parameters in total, group normalisation [11] throughout—batch statistics would be estimated on the public cohorts and applied elsewhere. Training uses AdamW ($\eta = 3 \times 10^{-4}$, weight decay 10^{-4}, batch 8), one-cycle scheduling over 40 epochs, bfloat16 autocast and weight EMA (decay 0.995); deployment averages five seeds with four-fold flip test-time augmentation. The challenge publishes no reference baseline, so variant A serves as one; we report no backbone comparison, since the probe answers the ranking question more directly.

4.2 Ablation Under Both Protocols

Table 1 reports the ablation ladder under both protocols, and the contrast is the point: scored on all 590 cases the three variants span 0.012 and the zero-pixel probe sits 0.035 below the best, whereas under masked-only evaluation they span 0.46 and the baseline is revealed to have no ordinal agreement. Dense supervision with structured pooling accounts for almost all of the recovered skill. Each row is one cross-validation run, which the next paragraph shows is not by itself enough to separate close variants.

A second measurement bounds conclusions more tightly than the intervals do. Re-running the identical pipeline—same folds, same seed, same schedule—moves the masked-only composite by up to 0.20: training nondeterminism amplified by the seven anchor cases, so a run happening to recover one LR-1 gains more than any architectural change we tested. **Differences below that floor cannot be attributed to a configuration on the evidence of single runs.**

Variant A survives it: its interval $[0.083, 0.098]$ is disjoint from both others, and the gaps to B and C—0.46 and 0.44—are more than twice the floor. B and C differ by only 0.02, well inside the 0.20 reproducibility floor established above, so this single-run gap cannot be trusted and we do not select between them on the primary number. Two criteria fixed in advance separate them. First, under the nested decision rule of Sect. 3.4, fitted inside each training fold and applied to the held-out fold, C reaches 0.555 against B's 0.493: B's rule does not survive leaving the data it was tuned on. Second, a transfer test holding out the 2.5 mm cohort ($n = 190$), averaged over three seeds, gives 0.576 ± 0.069 (A), 0.556 ± 0.046 (B) and $\mathbf{0.601 \pm 0.026}$ (C). C is both best and most stable under repetition—what matters when deployment is a single draw from one unseen institution. Single runs on this corpus are not evidence. Carried through both criteria, the zero-pixel probe reaches 0.380 nested—below B's 0.493, C's 0.555—and 0.053 ± 0.000 over three seeds on the 2.5 mm transfer ($n = 188$ masked), far below all three variants: the images' margin over the legitimate scalar signal is real and large, not just at the argmax.

Table 1. One set of cohort-grouped out-of-fold predictions, scored on all 590 cases as the official metric does and on the 521 masked cases as our deployment-condition diagnostic does, at the *argmax*: the rule of Sect. 3.4 is not applied. Intervals are 95% bootstrap (2000 resamples). The probe reads only spacing, thickness and diameter. A: baseline; B: adds dense supervision and structured pooling; C: adds phase differences.

	Scored on all 590		Scored on masked 521	
Variant	S	κ_{adj}	S (95% CI)	κ_{adj}
zero-pixel probe	0.871	0.932	0.459	0.450
A : baseline	0.894	0.943	0.090 [0.083, 0.098]	0.000
B : + dense sup., pooling	0.904	0.952	**0.553** [0.334, 0.686]	0.542
C : + phase differences	**0.906**	**0.955**	0.530 [0.300, 0.686]	0.514

4.3 Feature Extraction

Applying the LI-RADS table to *ground-truth* features yields $\kappa_{\text{adj}} = 0.922$: the ordinal problem is determined by three binary features and a diameter, so the real question is how accurately they can be extracted. Held-out cohort-grouped areas under the ROC curve move from the baseline's fold range to variant C thus: non-rim APHE .69–.80 → 0.929, rim APHE .46–.62 → 0.795, washout .62–.79 → 0.872, capsule .52–.72 → 0.757. Rim APHE moves from near-chance to usable—the mechanism behind Table 1.

4.4 Private Test Set

On the 174 private cases our entry scores $S = 0.409$ ($\kappa_{\text{adj}} = 0.40$, SCR = 0.46), fourth of six teams as of 17 August 2026; the leader scores 0.48 (0.47, 0.52), ahead on both components, and third place goes to an entry with a *lower* κ_{adj} than ours but far stronger SCR—the trade Sect. 5 anatomises. Our masked-only estimate of $\kappa_{\text{adj}} = 0.51$ overstates the realised 0.40 but is of the same order and within the run-to-run spread of Sect. 4.2, whereas the public-protocol estimate of 0.955 is not—the external evidence for Sect. 2.3. The submitted system is exactly that of Sect. 3, A and B being ablations rather than separate entries, and all selection was completed on internal cross-validation before that *single* submission.

5 Error Analysis and Failure Modes

Ordinal errors respect the scale. Table 2 gives the confusion matrix on the 521 masked cases. Of the 258 cases where both truth and prediction are ordinal, 82.9% are exact, 14.0% off by one and only 3.1% by two or more; of the errors alone, 82% are adjacent. The model has learned the ordering rather than behaving as a flat five-way classifier, and the largest confusion, LR-4 → LR-5 (22 cases), is adjacent and therefore cheap.

The special-category deficit is concentrated and explicable. LR-TIV recall is 0.36: 25 of 64 are predicted LR-M. Both being special categories, this costs

Table 2. Confusion matrix, masked-only cross-validation ($n = 521$). Rows are ground truth, columns predictions, diagonal in bold. The boxed cell dominates the SCR deficit.

	LR-1	LR-2	LR-3	LR-4	LR-5	LR-M	LR-TIV
LR-1 (4)	**0**	2	0	0	0	2	0
LR-2 (3)	0	**0**	1	0	1	1	0
LR-3 (16)	0	0	**0**	1	3	7	5
LR-4 (47)	1	0	3	**8**	22	10	3
LR-5 (258)	0	0	3	7	**206**	27	15
LR-M (129)	1	0	2	4	13	**91**	18
LR-TIV (64)	1	0	1	0	14	25	**23**

SCR directly while costing κ_{adj} nothing—exactly the shape of our leaderboard result. Comparing the 23 recovered against the 25 lost: the lost cases have median diameter 70 mm against 131 mm, mean P(LR-TIV) 0.172 against 0.739, and mean predicted rim-APHE 0.452—above the 0.328 of true LR-M. The model is confidently wrong, not uncertain: small-volume tumour in vein with rim enhancement reads as LR-M, coherent since limited venous invasion need not show the bulky intravascular mass defining it.

The annotation schema compounds this: APHE is unassessed for all 64 LR-TIV cases, so the head is trained where rim enhancement is a *perfect* indicator of LR-M—all 69 annotated rim cases—and never co-occurs with LR-TIV. Its confident rim prediction on small LR-TIV lesions is extrapolation beyond its supervision. The remedy is an annotation change, not architecture.

Accuracy is U-shaped in lesion size—55.0% exact below 20 mm (n=20), 71.0% at 20–40 mm (n=124), 66.0% at 40–80 mm (n=244), 51.1% above 80 mm (n=133)—too few voxels at the small end, infiltrative overlap of LR-5, LR-M and LR-TIV at the large end.

6 Discussion and Limitations

Why the method works, and where it cannot yet be checked. Applying the LI-RADS decision rules to ground-truth features reproduces the label at $\kappa_{\text{adj}} = 0.922$, so better feature extraction—not capacity, since the three variants lie within 0.1 M parameters—is the direct route to a better category, and dense supervision plus structured pooling delivers it. But special-category transfer cannot be validated on public data: eighty-nine per cent of LR-M and 83% of LR-TIV observations come from the 0.8 mm cohort, so holding it out leaves 14 LR-M cases in training and the fold measures class starvation, not domain shift—SCR was never measurable beforehand, and it is exactly where the private set penalised us.

Our protocol constrains the benign end weakly. Excluding mask-absent cases leaves seven benign observations—four LR-1, three LR-2—among the 328 ordinal-truth cases, of which 258 survive into κ_{adj}; LR-3 and LR-4 recall are near

zero and 0.17, a data limitation rather than an architectural one. A benchmark retaining lesion-free examinations *and* supplying masks would resolve it.

Two negative results. Training only on masked cases improves the argmax estimate (0.575 versus 0.530) but loses nested (0.548 versus 0.555), since mask-absent cases still carry usable signal for the benign class. External data also hurt: mapping the closest available match, MCT-LTDiag [9] (95 masked haemangiomas, 104 non-HCC malignancies), onto our absent benign category cut the masked-only score from 0.530 to 0.223, from rare LR-5 to LR-1 confusions and a washout-prevalence mismatch (49% internal versus 5% external). No pre-trained weights or external data were used anywhere.

Next steps. Against the failure of Sect. 5 we will test a vessel-aligned context branch and a size-conditioned LR-TIV/LR-M decision, but both residual weaknesses are bounded by the annotation schema more than the architecture: the benign end needs lesion-free cases carrying masks, the LR-TIV/LR-M boundary needs APHE assessed on tumour-in-vein.

Reproducibility. Inference runs in the official container (`codalab/codalab-legacy:gpu310`; Python 3.10, PyTorch 2.3.1, NumPy 1.26.4, CUDA 12.1), network disabled, one bundled package beyond the base image. Verified inside that image before submission: 18.1 s per case, 52.5 min for 174 cases against a three-hour limit, peak GPU memory 0.54 GB; training ran separately on one NVIDIA GB10 GPU at 47 min per fold. Splits are seeded (seed 42) and every number comes from the official evaluation script; code, weights and the container recipe will be released on publication. All data are the organisers' harmonisation of four public, de-identified retrospective collections [1,3,7,10]; no new patient data were collected.

7 Conclusion

Model selection on the public AMPLIFAI corpus is confounded by two independent shortcuts: acquisition leaks the special categories, and a lesion diameter that is zero exactly when the segmentation is absent leaks the ordinal scale—letting a classifier reading no voxel intensities score 0.871, while a baseline with no ordinal agreement scores $\kappa_{\text{adj}} = 0.943$. A cohort-grouped, masked-only protocol addresses both—fully for mask absence, only approximately for acquisition, whose dominant signature is split rather than held out—and under it LI-RADS-motivated components transfer to the private set, with the residual deficit localised to small-volume LR-TIV read as LR-M. Neither shortcut is specific to this dataset; both follow from harmonising independently collected cohorts, and the checks that expose them—a zero-pixel probe over acquisition metadata, and evaluation restricted to the annotations deployment guarantees—are inexpensive enough to run on any harmonised benchmark.

Acknowledgments. We thank the AMPLIFAI organisers for curating and releasing the dataset and evaluation code.

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

References

1. Bartnik, K., Bartczak, T., Krzyziński, M., et al.: WAW-TACE: a hepatocellular carcinoma multiphase CT dataset with segmentations, radiomics features, and clinical data. Radiol. Artif. Intell. **6**(6), e240296 (2024). https://doi.org/10.1148/ryai.240296
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). https://doi.org/10.1148/radiol.2018181494
3. Erickson, B., Kirk, S., Lee, Y., et al.: The Cancer Genome Atlas liver hepatocellular carcinoma collection (TCGA-LIHC). The Cancer Imaging Archive (2016). https://doi.org/10.7937/K9/TCIA.2016.IMMQW8UQ
4. Geirhos, R., Jacobsen, J.-H., Michaelis, C., et al.: Shortcut learning in deep neural networks. Nat. Mach. Intell. **2**(11), 665–673 (2020). https://doi.org/10.1038/s42256-020-00257-z
5. Kulkarni, P., Shah, N., Suryavanshi, A., et al.: 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
6. Llovet, J.M., Kelley, R.K., Villanueva, A., et al.: Hepatocellular carcinoma. Nat. Rev. Dis. Primers **7**(1), 6 (2021). https://doi.org/10.1038/s41572-020-00240-3
7. Luo, J., Wan, X., Du, J., et al.: Comprehensive multi-phase 3D contrast-enhanced CT imaging for primary liver cancer. Sci. Data **12**(1), 768 (2025). https://doi.org/10.1038/s41597-025-05125-2
8. Maier-Hein, L., Eisenmann, M., Reinke, A., et al.: Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun. **9**, 5217 (2018). https://doi.org/10.1038/s41467-018-07619-7
9. MCT-LTDiag: a multi-phase CT dataset for automated differential diagnosis of liver tumours. Harvard Dataverse (2025). https://doi.org/10.7910/DVN/S3RW15
10. Moawad, A.W., Morshid, A., Khalaf, A.M., et al.: Multimodality annotated hepatocellular carcinoma data set with imaging segmentation. Sci. Data **10**(1), 33 (2023). https://doi.org/10.1038/s41597-023-01928-3
11. Wu, Y., He, K.: Group normalization. In: ECCV 2018. LNCS, vol. 11217, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_1
12. Zech, J.R., Badgeley, M.A., Liu, M., et al.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLOS Med. **15**(11), e1002683 (2018). https://doi.org/10.1371/journal.pmed.1002683