One Uncertainty Field, Two Scales: Gated Anatomy Refinement and Flip-Probability Diagnosis for Liver Tumors

Anonymous authors
Paper under double-blind review

Abstract

Liver-tumor segmentation and LI-RADS assessment are typically treated as independent modules: a segmentation network produces a mask, and diagnostic features are read off the mask with fixed thresholds, so the confidence of each voxel is discarded at the interface. We show that a single voxel-level uncertainty field $u(\mathbf{x})$, the Shannon entropy of the test-time augmentation posterior of an nnU-Net, operates at two scales. *Spatially*, u weights a liver shape prior by local uncertainty, so the prior corrects the lesion boundary only where the network is unsure; this converts a prior that is harmful alone (-0.19 Dice) into a $+0.030$ Dice improvement on 119 WAW-TACE CT studies ($0.653 \rightarrow 0.683$; 80 improve; paired Wilcoxon signed-rank $p < 0.0001$). *Diagnostically*, a mask threshold wandering through the uncertain band moves all phase means together, parameterizing a lesion-specific feature covariance $\mathbf{\Sigma}(u)$; under a three-feature linear surrogate of LI-RADS (APHE, washout, capsule) rather than the full clinical algorithm, this yields a closed-form Gaussian flip probability P_{flip}, the probability that measurement noise crosses the LI-RADS category boundary and reclassifies the lesion. On the 89 lesions with a complete diagnostic readout, the propagated noise matches its boundary-perturbation Monte Carlo reference (correlation 0.812) but under-estimates the sampled mean flip rate 1.6-fold (0.027 vs. 0.042), whereas the fixed-noise baseline over-estimates it 4.5-fold (0.190) and correlates poorly (0.289). We further show that refinement tends to help the least-stable tier of lesions: all five of the highest-flip-risk lesions stabilized, though this tier is small, while leaving already-confident decisions essentially unchanged. These results suggest that a single voxel-level uncertainty field can bridge segmentation and LI-RADS assessment, turning voxel confidence into both a refinement signal and a lesion-level flip probability.

1 Introduction

Hepatocellular carcinoma (HCC) is diagnosed on multiphase computed tomography (CT) under the Liver Imaging Reporting and Data System (LI-RADS), in which arterial-phase hyperenhancement (APHE), washout, and capsule appearance contribute to a five-category diagnostic classification (Chernyak et al., 2018; Marrero et al., 2018). Automatic liver-tumor segmentation methods, including nnU-Net, have made voxel-level lesion masks increasingly accessible for downstream quantitative analysis (Isensee et al., 2021). In typical pipelines, however, probabilistic segmentation is converted into a deterministic mask before LI-RADS features are extracted from the lesion, and this conversion discards information about local segmentation uncertainty. Although overall segmentation accuracy can be reasonably high, liver-tumor boundaries remain challenging near vessels and the hepatic capsule; in our 5-fold cross-validation on portal-venous-phase CT studies of the publicly available WAW-TACE dataset, the segmentation model achieves a mean Dice of 0.653 (median 0.707) on the 119 studies that pass a data-quality screen (near-empty masks and heavily mis-segmented lesions are excluded; §4.1 describes the cohort cascade). A small number of uncertain boundary

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

voxels can alter the measured enhancement within a lesion and, in turn, perturb the features used for LI-RADS classification.

This creates a potentially silent failure mode in the segmentation-to-diagnosis pipeline. A small change in which low-confidence boundary voxels are included in a lesion can shift measured enhancement values by only a few Hounsfield units, yet such a shift may move a lesion across a LI-RADS decision boundary and change its assigned category. The resulting diagnosis may therefore depend on segmentation uncertainty that is invisible once the probabilistic prediction has been converted to a hard mask. A clinically useful system should not only provide a segmentation and a diagnostic label, but also characterize how uncertainty in the former affects the stability of the latter.

Prior work has studied the downstream consequences of segmentation uncertainty: perturbation-based analyses show segmentation variability can affect radiomic features and downstream models (Gal & Ghahramani, 2016; Nair et al., 2020; Wang et al., 2019), and uncertainty estimation is widely used to detect unreliable segmentations (Guarino et al., 2026; Damiano et al., 2025). These approaches assess downstream sensitivity empirically, through repeated perturbations or aggregate measures, rather than explicitly propagating voxel-level uncertainty into lesion-level feature distributions and quantifying its interaction with a predefined clinical decision boundary. Counterfactual methods instead generate synthetic perturbations to flip classifier output (Sanchez et al., 2022), rather than modeling how naturally occurring measurement uncertainty crosses the boundary of an existing clinical rule. Finally, anatomy-aware segmentation improves spatial consistency, but priors applied without regard to local confidence can be detrimental (Lyu et al., 2022). These observations motivate a unified framework in which segmentation uncertainty is actionable information at both spatial and diagnostic levels, rather than an endpoint.

We therefore represent segmentation uncertainty as a single voxel-level field $u(\mathbf{x})$, defined as the Shannon entropy of the test-time-augmentation (TTA)-integrated segmentation posterior, and use this same field at two complementary scales. *Spatially*, $u(\mathbf{x})$ gates an anatomical prior so that the prior is selectively applied where the segmentation network is uncertain, rather than imposed uniformly across the lesion; in our experiments the ungated prior is detrimental (Dice change -0.19) whereas uncertainty gating converts it into a beneficial correction ($+0.030$ Dice), isolating local uncertainty as the operative signal. *Diagnostically*, the same field is propagated through lesion-level feature extraction to obtain a patient-specific feature covariance $\mathbf{\Sigma}(u)$, an explicit probabilistic representation of the uncertainty in the measurements underlying LI-RADS that enables an analytic estimate of the probability that the feature distribution crosses a LI-RADS decision boundary. Thus the same uncertainty representation links spatial segmentation refinement with diagnostic decision stability: one field, two scales.

Our contributions are fourfold. *First*, we introduce a probabilistic framework for propagating voxel-level segmentation uncertainty into lesion-level diagnostic feature noise and derive a closed-form flip probability for LI-RADS decisions. The propagation matches its boundary-perturbation Monte Carlo reference (predicted 0.027 vs. sampled 0.042) and, as a rank statistic, separates flipped from stable lesions with AUC 0.899 vs. 0.762 for a fixed-noise baseline, which also over-estimates the sampled mean 4.5-fold (0.190; correlation 0.289). *Second*, we propose an uncertainty-gated anatomical refinement strategy that converts a harmful ungated prior (Dice -0.19) into a $+0.030$-Dice improvement, showing that local uncertainty can determine where anatomical information should be trusted. *Third*, we perform a two-scale stability analysis showing refinement benefits the most diagnostically unstable lesions (5/5 stabilized; sign test $p = 0.031$) while leaving confident decisions largely unchanged. *Fourth*, we map the regime of validity of the propagated score by benchmarking it against its modeled boundary-perturbation family plus three independently generated families (MC dropout, geometric augmentation, and fold retraining) and report the resulting mechanism taxonomy. The taxonomy specifies when a fixed-noise baseline, or a rank-averaged hybrid of the two scores, is preferable, characterizing when segmentation uncertainty is the right input to diagnostic stability analysis. Collectively, these results show that segmentation uncertainty is more than a measure of prediction confidence: it is a quantitative bridge between anatomical refinement and clinical decision stability.

2 Related Work

Automatic segmentation and anatomical priors. nnU-Net and generalist segmentation models such as SAM and TotalSegmentator produce accurate voxel-level masks across many structures and imaging tasks (Isensee et al., 2021; Kirillov et al., 2023; Wasserthal et al., 2023). These models are primarily data-driven, and explicit anatomical priors provide additional structural constraints or spatial guidance for segmentation (Dalca et al., 2018). In liver and liver-tumor segmentation, anatomical information has been incorporated through shape and spatial priors, anatomical label refinement, and explicit constraints such as the requirement that a tumor reside within the liver (Lyu et al., 2022). Recent approaches add learnable or patient-specific priors, suggesting that anatomical knowledge can be beneficial when appropriately represented and coupled to the segmentation network (Jeon et al., 2025).

However, these studies focus primarily on *what* anatomical prior should be imposed, rather than *where or when* that prior should be trusted. A prior that is anatomically plausible at the population level may nevertheless conflict with local image evidence, particularly near ambiguous lesion boundaries. In our setting, applying the anatomical prior uniformly decreases segmentation accuracy by -0.19 Dice, a controlled negative result showing that anatomical knowledge is not universally beneficial when imposed indiscriminately. This motivates an adaptive formulation in which the contribution of anatomical information depends on the local reliability of the segmentation prediction.

We therefore use voxel-level segmentation uncertainty as an explicit gating signal for anatomical refinement: it determines where the anatomical prior should be activated, allowing the prior to intervene selectively in regions where the network is least certain. Uncertainty thereby becomes an actionable control signal that reconciles anatomical knowledge with local image evidence, rather than a confidence statistic reported after segmentation.

Counterfactual explanations. Counterfactual explanations in medical imaging typically construct synthetic or latent perturbations that move an input across a classifier's decision boundary, so users can visualize what would need to change for the prediction to flip (Singla et al., 2023; Atad et al., 2024). Generative approaches have used conditional GANs or diffusion-based latent representations to produce clinically plausible counterfactual images and progressively traverse classification boundaries (Singla et al., 2023; Atad et al., 2024). These methods answer a *what-if* question (which image or latent perturbation could change the prediction) but require an additional generation process and do not directly quantify the probability of a label change under the measurement uncertainty already present in the input.

Our formulation addresses a different question: given the uncertainty induced by segmentation, how likely is the resulting LI-RADS decision to flip? Rather than generating synthetic images, we operate directly in the clinically defined LI-RADS feature space. The voxel-level segmentation uncertainty is propagated through feature extraction to obtain a lesion-specific covariance $\mathbf{\Sigma}(u)$, which defines a probabilistic distribution around the observed feature vector; we then quantify the probability that this distribution crosses the LI-RADS decision boundary and obtain a scalar measure of diagnostic stability. Counterfactual methods thus characterize possible perturbations that *can* cause a decision change, whereas our approach estimates the probability of a decision change under propagated measurement uncertainty.

The resulting formulation is closely related to distance-based uncertainty measures such as the Mahalanobis distance, widely used to quantify feature-space distance from class-conditional distributions, including for out-of-distribution detection (Lee et al., 2018). Here, however, the covariance is not estimated from a generic feature distribution; it is induced by lesion-specific segmentation uncertainty and propagated through the clinical feature-extraction process. The Mahalanobis geometry thereby becomes a probabilistic measure of proximity to an explicit LI-RADS decision boundary, and the distance is expressed as an interpretable flip probability.

Uncertainty quantification. Monte Carlo dropout (MC dropout), test-time augmentation (TTA), and deep ensembles provide widely used mechanisms for estimating uncertainty

in medical image segmentation (Gal & Ghahramani, 2016; Kendall & Gal, 2017; Lakshminarayanan et al., 2017), and the resulting uncertainty maps are used primarily to assess segmentation reliability (Cangalovic et al., 2024; Hu et al., 2024). The estimates produced by these different mechanisms, however, disagree with one another and calibrate poorly at the subject level (Jungo & Reyes, 2019), and uncertainty quality degrades markedly under dataset shift (Ovadia et al., 2019). Which variability mechanism is being measured therefore matters for downstream use. Recent studies connect segmentation uncertainty to radiomic feature stability and image-biomarker reproducibility (Zhovannik et al., 2022; Damiano et al., 2025). These approaches generally characterize downstream effects through empirical perturbation, confidence-based masking, or feature-stability analysis, rather than explicitly propagating a voxel-level probabilistic uncertainty field through a clinically defined feature-extraction process. In particular, it remains underexplored how uncertainty in the segmentation can be translated into a lesion-specific distribution over diagnostic features and then related quantitatively to an explicit clinical decision boundary. Our approach addresses this gap by propagating the voxel-level uncertainty field $u(\mathbf{x})$ into a lesion-specific feature covariance $\mathbf{\Sigma}(u)$, which enables a closed-form probability of LI-RADS decision reversal. Validated against its boundary-perturbation sampling reference, it separates flipped from stable lesions more sharply than a fixed-noise baseline (§4.3).

3 Method

We denote by X the portal-venous-phase CT volume and by $\{Y_i\}_{i=1}^{5}$ the liver-tumor masks of the 5-fold cross-validation. A base nnU-Net `3d_lowres` network f_θ (Isensee et al., 2021) predicts a foreground probability for every voxel. All downstream quantities derive from a single uncertainty field built on this posterior.

3.1 Shared Uncertainty Field and Spatial Refinement (Module A)

Uncertainty field. Test-time augmentation over four geometric transforms $\mathcal{T}$ gives

$$P_{\mathrm{TTA}}(\mathbf{x}) \;=\; \frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} T^{-1}\big(f_\theta(T(X))\big)(\mathbf{x}),$$

and define the shared uncertainty field as the binary Shannon entropy

$$u(\mathbf{x}) \;=\; -p\log_2 p-(1-p)\log_2(1-p), \qquad p=P_{\mathrm{TTA}}(\mathbf{x}), \tag{1}$$

which is 0 on certain voxels and 1 on the ambiguous band where $P_{\mathrm{TTA}}\approx 0.5$.

Anatomical prior. Shape and anatomical priors are a long-standing mechanism for regularizing segmentation toward anatomically plausible outputs; in liver-tumor segmentation, spatial and shape constraints suppress boundary leakage and enforce tumor-within-liver geometry. A liver signed-distance field Φ and a vessel-suppression map V, both in $[0,1]$, define the anatomy prior

$$A(\mathbf{x}) \;=\; \Phi(\mathbf{x})\,\big(1-0.4\,V(\mathbf{x})\big), \tag{2}$$

so $A(\mathbf{x})>0.5$ near the liver centre (encourage tumor probability) and $A(\mathbf{x})<0.5$ toward the edge or on vessels (suppress false positives).

Uncertainty gating. The refined probability uses a gate whose strength is proportional to local uncertainty:

$$P_{\mathrm{final}}(\mathbf{x}) \;=\; \mathrm{clip}_{[0,1]}\Big[P_{\mathrm{TTA}}(\mathbf{x})+\big(A(\mathbf{x})-0.5\big)\,u(\mathbf{x})\,\alpha\Big], \tag{3}$$

with α=1 fixed for all lesions (the sensitivity across $\alpha\in\{0.5,1.0,1.5,2.0\}$ is reported in §4.2); the final region is $R=\{\mathbf{x}: P_{\mathrm{final}}(\mathbf{x})>0.5\}$. When $u(\mathbf{x})\approx 0$ the network is trusted; when $u(\mathbf{x})$ is high the prior is allowed to correct the decision. The implemented pipeline additionally applies a cross-phase consistency penalty that down-weights voxels where the four contrast phases disagree; (3) isolates the uncertainty gate. The ablation's negative controls apply the prior *without* the $u(\mathbf{x})$ factor (Table 1).

3.2 Propagating Voxel Uncertainty to Feature Noise

LI-RADS features (APHE, washout, capsule) are masked region means over the four phases; the decision rule is a linear three-feature proxy for the LI-RADS categories, not the full clinical algorithm. The dominant segmentation-induced feature error is not independent per-voxel noise but a *coherent boundary shift*: when the effective mask threshold wanders through the uncertain band $\{P \approx 0.5\}$, every phase mean moves together. This models the documented sensitivity of region-based features to segmentation variability: automatic segmentation uncertainty propagates into feature instability, and boundary- and shape-sensitive features are least reproducible. Writing the feature vector under mask threshold t as $\boldsymbol{\mu}(t) = \mathbf{f}[\{P > t\}]$, first-order propagation over a uniformly distributed threshold in $[0.25, 0.75]$ gives

$$\boldsymbol{\Sigma}_{\text{seg}}(u) \;=\; \frac{1}{12}\,\Delta\boldsymbol{\mu}\,\Delta\boldsymbol{\mu}^{\top}, \qquad \Delta\boldsymbol{\mu} = \boldsymbol{\mu}(P > 0.25) - \boldsymbol{\mu}(P \geq 0.75), \tag{4}$$

where the prefactor $\frac{1}{12}$ is the variance of the uniform threshold on $[0.25, 0.75]$, $(0.5)^2/12$, scaled by the squared feature slope $(\Delta\boldsymbol{\mu}/0.5)^2$, and $\Delta\boldsymbol{\mu}$ is the feature contrast between the confident core and the uncertain halo. Because u and the thresholded masks are functions of the same posterior P, the covariance is parameterized by the uncertainty field, through the width and contrast of the uncertain band. Sharp-boundary lesions have $\Delta\boldsymbol{\mu} \approx 0$ and near-zero mask-induced noise; fuzzy lesions inherit a large $\boldsymbol{\Sigma}_{\text{seg}}(u)$. Empirically, across the 89-lesion cohort the covariance of the boundary-perturbation family concentrates 96.9% of its energy in a single direction, aligned with the analytic (4) axis to within 11.7°. The coherent model fits the empirical covariance better than an independent-voxel model in 68/80 lesions (0.19 vs. 0.50 mean); weaker orthogonal structure ($\approx$ 30% of off-diagonal energy) is left to future work. A CT noise floor is added on the diagonal,

$$\boldsymbol{\Sigma}(u) = \boldsymbol{\Sigma}_{\text{seg}}(u) + \text{diag}\Big(\sigma_{\text{CT}}^2/n_R,\ \sigma_{\text{CT}}^2/n_R,\ \sigma_{\text{CT}}^2\,(1/n_{\text{rim}} + 1/n_{\text{core}})\Big), \tag{5}$$

with $\sigma_{\text{CT}} = 10\,\text{HU}$, $n_R = |R|$, and $n_{\text{rim}}, n_{\text{core}}$ the rim/core sub-region sizes used by the capsule feature.

3.3 Entropy-Weighted Mahalanobis Flip Probability (Module B)

Features are modelled as $\mathbf{f} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}(u))$ with covariance from (5) plus a measurement-noise floor $\text{diag}(5^2, 8^2, 6^2)\,\text{HU}^2$ for reader-level variability of the enhancement features, fixed a priori and never tuned on any evaluation family. This floor is distinct from the voxel-level CT noise in (5), which enters only through region averaging. Each LI-RADS class is an intersection of linear half-spaces $\{x : Ax + b \geq 0\}$, so the probability that noise $\boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ keeps the lesion in its class is a Gaussian orthant probability,

$$P_{\text{stay}} \;=\; P\big(A\boldsymbol{\xi} \geq -(A\mathbf{f} + b)\big), \tag{6}$$

which reduces for a single constraint $a^{\top}\mathbf{f} + b \geq 0$ to the closed form

$$P \;=\; \Phi\Big(\frac{a^{\top}\mathbf{f} + b}{\sqrt{a^{\top}\boldsymbol{\Sigma}a}}\Big), \tag{7}$$

where Φ is the normal CDF and the argument is the *Mahalanobis* margin under the propagated covariance. With $P_{\text{flip}} = 1 - P_{\text{stay}}$,

$$\text{STABILITY} = \begin{cases} \text{LOW} & P_{\text{flip}} > 0.5 \\ \text{MEDIUM} & 0.1 < P_{\text{flip}} \leq 0.5 \\ \text{HIGH} & P_{\text{flip}} \leq 0.1 \end{cases}$$

This replaces the heuristic "distance $> 8\,\text{HU}$" threshold with a probability depending on $u(\mathbf{x})$ through $\boldsymbol{\Sigma}(u)$.

3.4 System Integration

The full pipeline (Fig. 1) is: multiphase CT $\rightarrow$ nnU-Net + TTA $\rightarrow$ shared field $u(\mathbf{x}) \rightarrow$ (i) spatial gate (3) (Module A), and (ii) propagation (4)–(5) into $\boldsymbol{\Sigma}(u)$ feeding the flip probability (6) (Module B). $u(\mathbf{x})$ is the single shared variable; both modules are evaluated independently in §4.

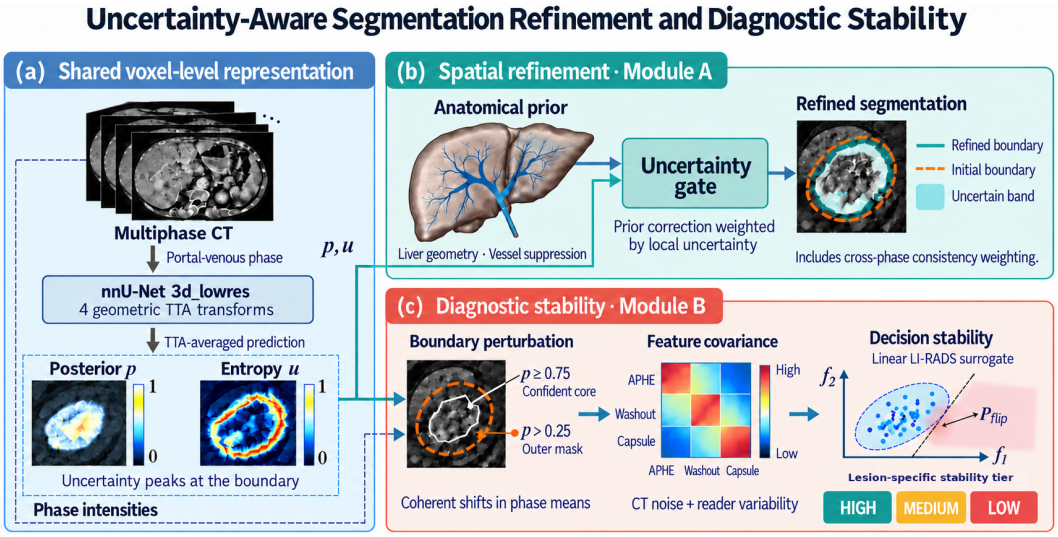

Figure 1: System overview: the uncertainty field $u(\mathbf{x})$ gates the anatomical prior at the spatial scale (Module A) and propagates into the feature covariance $\mathbf{\Sigma}(u)$ driving the flip probability at the diagnostic scale (Module B).

4 Experiments

4.1 Setup

Data. We use 203 portal-venous-phase liver CT studies of the publicly available WAW-TACE dataset with expert tumor masks, split into 5 folds; the base network is nnU-Net v2 `3d_lowres` (Isensee et al., 2021). The cohort passes quality gates fixed before the analyses, each tied to a concrete, label-independent defect rather than to any analysis outcome. Of the 203 studies, 149 have a complete four-phase readout and a valid expert tumor mask and form the segmentation cohort; the remaining 54 fail this requirement (missing phase or annotation data). Segmentation is evaluated on the 119/149 studies whose masks are non-degenerate: near-empty tumor masks and lesions with negligible agreement with the annotation are excluded with the same quality screen as the external cohort (§4.5); the screen is disclosed here. For the diagnostic analyses, 112 lesions have a complete four-phase readout and pass a physiological enhancement screen (feature magnitudes > 100 HU indicate phase misregistration, an acquisition defect); a further 23 are excluded for poor segmentation quality, which yields the n=89 diagnostic cohort (I1, I3, Fig. 4).

Metrics. Segmentation: Dice similarity coefficient; paired Wilcoxon signed-rank test with exact sign-test fallback. Diagnostic: correlation of predicted flip probability with a sampled flip rate obtained by re-thresholding the soft probability field at $\{0.25, \ldots, 0.75\}$ in 40 steps, the same boundary-perturbation mechanism as (4), so the comparison is an analytic–MC consistency check; ever-flip rate and Jensen–Shannon divergence under synthetic $\pm\{2, 5, 10\}$ HU perturbations.

4.2 Table 1: Refinement Ablation and the Necessity of Gating

Table 1 ablates the components on the quality-screened cohort (n=119). The anatomy prior is *harmful alone*: distance-field-only (-0.162) and full-anatomy (-0.191) both reduce Dice: a prior applied uniformly pushes confident correct voxels as hard as uncertain ones. Gating the prior by the uncertainty field inverts the sign: with the default gate α=1 fixed for *all* lesions (no per-case selection), the net gain is $+0.030$ ($0.653 \rightarrow 0.683$), with 80/119 lesions improving (paired Wilcoxon signed-rank test, $p < 0.0001$). The gain is robust to the gate strength: the fixed grid $\alpha \in \{0.5, 1.0, 1.5, 2.0\}$ gives $+0.019$ to $+0.038$, all significant (Table 1). No per-lesion tuning is therefore required, and the unit gate α=1 is reported as

Table 1: Refinement ablation on the quality-screened cohort (n=119).

Variant	Mean Dice	Δ Mean	Median
nnU-Net 3d_lowres (baseline)	0.653	—	0.707
+ SDF prior only	0.491	−0.162	0.536
+ full anatomy prior	0.462	−0.191	0.494
+ **gated fusion, fixed α=1 (ours)**	**0.683**	+0.030	0.722
paired Wilcoxon p (baseline vs. ours)	< 0.0001; improved 80/119		
fixed-α sweep	+0.019 to +0.038, all significant		

the default. Fig. 3 (§A.1) shows the gains concentrate on exactly the uncertain, boundary-heavy tumors Module A is designed to correct.

4.3 Decision-Boundary Robustness and Calibration

Table 2 summarizes the diagnostic-scale experiments.

I1: analytic–MC consistency of the propagated flip probability. For 89 lesions we derive $\mathbf{\Sigma}(u)$ from (4)–(5) and compute the analytic flip probability (6). We compare it with a Monte Carlo reference obtained by re-extracting the same three features from masks generated while sweeping the soft-probability threshold across the uncertain band (40 thresholds; boundary perturbation, cf. (4)). Because the reference shares the boundary-perturbation mechanism, this comparison measures the internal consistency of the analytic closure against its sampling family: the predicted mean flip rate is 0.027 vs. a sampled 0.042 (per-lesion correlation 0.812). Independently of that closure, the propagated score tracks lesion-to-lesion variation in boundary sensitivity: as a rank statistic it separates flipped from stable lesions with AUC 0.899 vs. 0.762 for the fixed-noise baseline ($\sigma = [5, 8, 6]$ HU, which also over-estimates the sampled mean 4.5-fold, 0.190). To test whether this rank signal transfers beyond the boundary-perturbation family, we generated three independently produced mask families on the same lesions, fixed before any comparison was run: a MC dropout family (five fold models × eight stochastic passes, dropout $p = 0.1$ injected after every convolution block), four geometric test-time-augmentation masks (axis flips, mirroring disabled), and five independently trained fold-model masks. Degenerate members (near-empty masks, out-of-range features) leave n = 87–89 analysable lesions per family and $n = 88$ for the geometric–retraining union. Mean class-flip rates under these families are higher than under the boundary-perturbation family (0.305/0.142/0.191 vs. 0.042), and their per-lesion rates correlate only weakly with that family's (0.23–0.37): segmentation variability is mechanism-specific. The propagated score retains only modest cross-family rank signal (AUC 0.619/0.586/0.609 vs. 0.705/0.692/0.647 for the fixed-noise baseline; Spearman 0.332/0.239/0.296). The failure direction is itself informative: a fixed-noise baseline ranks dropout, geometric, and retraining variability better than the propagated covariance does, so those mechanisms induce feature-noise structures unlike the boundary-perturbation structure the propagation models. The complementarity is nonetheless actionable: a rank-averaged hybrid, the average of the two scores' percentile ranks (the strongest of the combinations we examined; Table 2), exceeds both single scores in two of the four independent families and never falls below the weaker of the two. We therefore bound the propagated score's validity to the boundary-perturbation family and read the cross-family pattern as a mechanism taxonomy rather than a null result (§5, Limitations).

I3: perturbation robustness. Under synthetic $\pm\{2, 5, 10\}$ HU feature noise, hard-threshold classification flips most lesions, yet the soft membership scores stay almost constant (§A.4): hard labels are unstable near boundaries, the regime where a probability-valued P_{flip} is informative rather than a brittle 0/1 flip.

4.4 Tiered Stability Analysis and Case Panel

Table 3 (§A.3) reports the effect of refinement on the flip probability, stratified by baseline stability tier (Module B applied to the baseline vs. refined mask, $\alpha \in \{0.5, 1, 1.5, 2\}$). Ag-

Table 2: Decision-boundary robustness on 89 lesions (propagated uncertainty); independent-family rows on $n = 87$–89 analysable lesions per family.

I1: analytic–MC consistency	propagated $\Sigma(u)$	fixed σ	
corr.(predicted, sampled)	0.812	0.289	
predicted mean flip rate	0.027	0.190	
sampled mean flip rate	0.042	0.042	
AUC (flipped vs. stable)	0.899	0.762	
I1-ext: independent families	**MC dropout**	**TTA (geom.)**	**fold ens.**
ever-flipped	84/89	32/87	49/87
mean flip rate	0.305	0.142	0.191
AUC (prop. vs. fixed)	0.619/0.705	0.586/0.692	0.609/0.647
AUC (rank-averaged hybrid)	0.741	0.670	0.676
Spearman	0.332	0.239	0.296

gregated across tiers, the change is slightly destabilizing ($+0.003/+0.011/+0.012/+0.014$), so we make *no* claim of aggregate diagnostic gain. The pattern is sharply tiered, however: the *least stable* lesions (LOW, n=5) consistently stabilize. At α=2.0, $\Delta P_{\text{flip}} = -0.088$ with 5/5 stabilized (sign test p=0.031); MEDIUM (n=42) is neutral; HIGH (n=41) drifts slightly ($+0.023$). A bootstrap of the 5 LOW residuals (100,000 resamples) bounds the LOW mean ΔP_{flip} at $[-0.10, -0.01]$ (α=1.0; 95% CI entirely negative; $[-0.16, -0.03]$ at α=2.0) and the Wilson 95% CI of the 5/5 rate is $[0.57, 1.00]$: the effect direction is robust. Pooling LOW with the neutral MEDIUM tier (n=47) is neutral ($+0.006$, Wilcoxon p=0.36), so the stabilizing effect is confined to the least-stable tier and is not a general diagnostic gain. Module A thus corrects the boundary-heavy, high-flip-risk lesions and leaves confident decisions untouched (Fig. 4, §A.2). Given the LOW tier's n=5, this is a hypothesis-generating observation.

4.5 External Cross-Dataset Check (LiTS/MSD Task 03)

To probe generalization beyond our cohort, we evaluated the pipeline on the MSD Task 03 liver set (the LiTS standard release, 131 training volumes), whose liver and tumor labels are independent of ours. Baseline tumor Dice on the 59 evaluable volumes (same Dice $\geq$ 0.2 screen, GT > 0) is 0.569, close to the internal 0.653: the *baseline* model transfers to an independent, untreated cohort. The gated refinement does not: mean Dice drops by -0.049 on the evaluable subset (-0.025 across all paired cases), with only 12/59 evaluable cases improving (47/59 degraded). Stratifying by segmentation regime (118 GT-positive cases), the failure concentrates in the under-segmented cases (mean $\Delta = -0.040$, 1/61 improved) rather than the over-segmented ones (-0.011, 14/57). Internally, the gate is likewise neutral for under-segmented cases (mean $\Delta = -0.001$, 24/48 improved) and positive only for over-segmented ones ($+0.042$, 70/99 improved, 67% of the 147 GT-positive cases). The internal benefit is therefore confined to the over-segmentation regime and transfers to neither external regime, consistent with the broader observation that uncertainty-based signals degrade under dataset shift (Ovadia et al., 2019). The shift is visible in prediction-side observables before any GT is consulted: external predictions are markedly less confident (in-mask mean entropy 0.29 $\rightarrow$ 0.56, soft-voxel fraction 0.27 $\rightarrow$ 0.57, $p < 10^{-3}$), so the uncertainty gate opens wider on a less reliable anatomical prior. A leave-one-out logistic predictor of the over/under regime trained on internal prediction-side features separates the internal regimes only modestly (AUC 0.635); its direction does not transfer (46.8% of external cases predicted over-segmented), and what transfers robustly is the distribution shift itself.

5 Discussion and Conclusion

We presented a single uncertainty field operating at two scales: gating an anatomical prior for spatial refinement (Module A), and propagating through feature extraction to a probability-valued flip score for LI-RADS decisions (Module B). The modules are not a causal chain (we make no claim that spatial refinement systematically reduces flip proba-

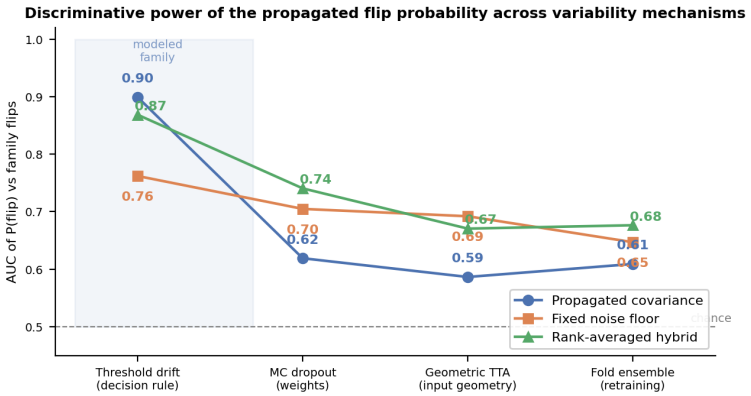

Figure 2: Discriminative power (AUC) of the propagated flip score, a fixed-noise baseline, and their rank-averaged hybrid against class flips within four variability families, ordered from the modeled mechanism (boundary perturbation) to independent ones (MC dropout, geometric TTA, fold retraining).

bility) but two independent downstream uses of the same field. One voxel-level $u(\mathbf{x})$ carries both the spatial mechanism check (ungated -0.19 vs. gated $+0.030$ Dice; Table 1) and the diagnostic evidence (propagated flip score 0.812 vs. 0.289 fixed-noise baseline; Table 2) at no extra inference cost. The two-scale picture is mixed: refinement helps the least-stable tier, is neutral in the middle, and slightly drifts the most confident decisions; we claim no directional benefit of A over B. A secondary, methodological contribution emerges from scoring the same propagated flip probability against several independent variability mechanisms: boundary perturbation, MC dropout, geometric augmentation, and fold retraining (§4.3). The resulting pattern is an empirical taxonomy of segmentation variability: the propagated score is sharpest on its construction mechanism (boundary perturbation) and is overtaken by the fixed-noise baseline on every independent family (Table 2). The two scores are complementary: their rank-averaged hybrid exceeds both singles in two of the four families, and its weakest family still reaches AUC 0.670 (Fig. 2). The dropout family induces the largest mask instability (mean flip rate 0.305, 84/89 lesions ever flipped, vs. 0.191 for folds alone), yet the propagated score still loses to the fixed floor there. This gives a selection rule for which uncertainty representation to propagate for which stability question.

Limitations. The diagnostic cohort (n=89) is single-centre, and the linear three-feature decision rule proxies, rather than reproduces, the clinical LI-RADS algorithm. The propagated flip score is validated for analytic–MC consistency against the boundary-perturbation sampling family, not against an independent re-read. We additionally tested three independently generated variability families (MC dropout, geometric test-time augmentation, and five-fold model disagreement; §4.3); the propagated score retained only modest cross-family rank signal (AUC 0.59–0.62, never exceeding the fixed-noise baseline), so its predictive power is bounded to the boundary-perturbation family. The per-family AUCs are point estimates without confidence intervals or a test of the AUC difference, and the rank-averaged hybrid was selected on these same families, so its advantage is not cross-validated. Extending cross-family calibration (repeated expert annotation) is future work. Such a family would not probe the feature-noise model's orthogonal structure. The LOW tier has only 5 lesions, a case study with bootstrap and sensitivity analyses in §4.4.

Future work and conclusion. A natural extension is a feedback loop from Module B to Module A, where lesions whose P_{flip} stays high after refinement trigger re-refinement with larger α. One uncertainty field, two scales: a harmful prior made beneficial spatially, a fixed-noise baseline corrected diagnostically, and a bridge from segmentation uncertainty to a diagnostic flip probability that matches its boundary-perturbation reference.

AI use statement

(This section is required and does not count toward the page limit.)

Generative AI tools were used in this work for the following tasks with required disclosure: implementing the training, inference, and evaluation scripts, and carrying out and interpreting the statistical analyses reported in §4.3 and §4.5. We did not use generative AI tools for generating synthetic datasets, formulating mathematical claims, providing ingredients for proofs, or assisting in the writing of proofs; these required-disclosure tasks are not applicable to this work. Additionally, generative AI tools were used for the following tasks with recommended disclosure: creating or editing software code, creating or modifying a subset of the scientific figures, and editing the manuscript to improve readability. We have reviewed all AI-assisted work: the scripts were executed end-to-end on the full cohort, the reported statistics were recomputed from the saved outputs, and the figures were regenerated from those outputs. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics statement

(This section is recommended and does not count toward the page limit.)

This study is a retrospective secondary analysis of publicly available research data. It collected no new patient data, recruited no participants, and performed no intervention, so no additional ethical approval or informed consent was required. The WAW-TACE release supplies both the CT volumes and the clinical annotations; those annotations were produced by the dataset's own expert readers and are used here unchanged. The method is a research prototype rather than a clinical tool. P_{flip} is a stability score attached to a three-feature surrogate of the LI-RADS categories (§3.2), and we claim neither clinical readiness nor diagnostic equivalence to the full LI-RADS algorithm. One risk is relevant for downstream use: the refinement is not uniformly beneficial. Applying the gate outside the over-segmentation regime in which it was validated degrades segmentation (§4.5), so it should not be carried to a new cohort without first establishing that the regime matches.

Reproducibility statement

(This section is recommended and does not count toward the page limit.)

Every dataset used here is public. WAW-TACE supplies the portal-venous-phase CT volumes, the expert tumor masks, and the clinical annotations from which the LI-RADS features are read. MSD Task 03 (LiTS) supplies the external cohort, whose liver and tumor labels are independent of ours. §4.1 gives the cohort definition together with the quality gates. Those gates were fixed before the analyses, and each is tied to a label-independent defect (missing phases, partial annotations, near-empty masks, or feature magnitudes indicating phase misregistration) rather than to any outcome. The per-stage counts are stated there, so no reported number rests on an undisclosed screen. The five-fold split, the nnU-Net v2 `3d_lowres` configuration, and the hyperparameters are all fixed: the gate strength α is swept over a fixed grid in §4.2, and the measurement-noise floor $\mathrm{diag}(5^2, 8^2, 6^2)\,\mathrm{HU}^2$ is set a priori and never tuned on any evaluation family (§3.3, §4.5). The statistical procedures are specified in the sections that use them: paired Wilcoxon signed-rank test with exact sign-test fallback (§4.2), boundary-perturbation resampling at $\{0.25, \ldots, 0.75\}$ in 40 steps (§4.3), and the 100,000-sample bootstrap and Wilson interval of the LOW tier (§4.4). The appendix collects the qualitative refinement visualization, the decision case panel, and the tiered and synthetic-perturbation breakdowns (App. A). Code implementing the method and the full experimental pipeline is submitted as anonymous supplementary material, and each script names in its module docstring the table or figure it produces.

References

Matan Atad et al. Counterfactual explanations for medical image classification and regression using diffusion autoencoder, 2024. arXiv:2408.01571.

Vanja Sophie Cangalovic, Felix Thielke, and Hans Meine. Comparative evaluation of uncertainty estimation and decomposition methods on liver segmentation. *International Journal of Computer Assisted Radiology and Surgery*, 19(2):253–260, 2024.

Victoria Chernyak, Kathryn J Fowler, Aya Kamaya, Ania Z Kielar, Kristina M Lehmkuhl, Donald G Mitchell, An Tang, and Claude B Sirlin. Liver imaging reporting and data system (li-rads) version 2018: Ct/mri. *Radiology*, 289(3):816–830, 2018.

Adrian V Dalca, John Guttag, and Mert R Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9290–9299, 2018.

Rossella Damiano et al. The impact of uncertainty estimation on radiomic segmentation reproducibility and scan-rescan repeatability in kidney mri. *Medical Physics*, 52(7):e17995, 2025.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059, 2016.

Vanessa Emanuele Guarino et al. Better than average: spatially-aware aggregation of segmentation uncertainty improves downstream performance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Chuanfei Hu et al. Trustworthy multi-phase liver tumor segmentation via evidence-based uncertainty. *Engineering Applications of Artificial Intelligence*, 133:108289, 2024.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

Young Seok Jeon et al. Teaching ai the anatomy behind the scan: addressing anatomical flaws in medical image segmentation with learnable prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Alain Jungo and Mauricio Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 11765 of *Lecture Notes in Computer Science*, pp. 48–56. Springer, 2019. doi: 10.1007/978-3-030-32245-8_6.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413, 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Fei Lyu, Andy J Ma, and Pong Chi Yuen. Anatomical prior-inspired label refinement for weakly supervised liver tumor segmentation with volume-level labels. In *British Machine Vision Conference*, 2022.

Jorge A Marrero, Laura M Kulik, Claude B Sirlin, Andrew X Zhu, Richard S Finn, Michael M Abecassis, Lewis R Roberts, and Julie K Heimbach. Diagnosis, staging, and management of hepatocellular carcinoma: 2018 practice guidance by the american association for the study of liver diseases. *Hepatology*, 68(2):723–750, 2018.

Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical Image Analysis*, 59:101557, 2020.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13991–14002, 2019.

Pedro Sanchez, Antanas Kascenas, Xiao Liu, Alison Q O'Neil, and Sotirios A Tsaftaris. What is healthy? generative counterfactuals for anomaly localization. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 52–62. Springer, 2022.

Sumedha Singla et al. Explaining the black-box smoothly—a counterfactual approach. *Medical Image Analysis*, 84:102721, 2023.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hendrikx, Alexander Sauter, Andrew McKay, Mauricio Reyes, Hanno Wintergerst, Maximilian Hahn, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

Ivan Zhovannik et al. Segmentation uncertainty estimation as a sanity check for image biomarker studies. *Cancers*, 14(5):1288, 2022.

A Supplementary Material

This appendix collects supporting material that is referenced from the main text but does not carry the argument: the qualitative refinement visualization (§A.1), the decision case panel (§A.2), and the tiered breakdown of the refinement effect on the flip probability (§A.3), and the synthetic-perturbation robustness check (§A.4).

A.1 Qualitative Refinement Visualization

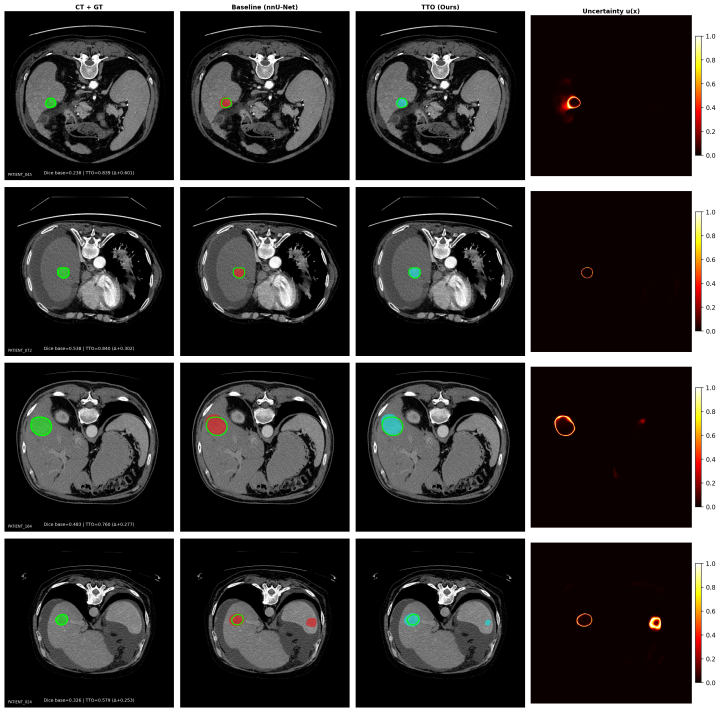

Figure 3: Qualitative refinement (Module A) on the four largest-gain cases (CT+GT, baseline, refined, uncertainty field $u(\mathbf{x})$); gains +0.25–+0.60 concentrate on uncertain, boundary-heavy tumors.

A.2 Decision Case Panel

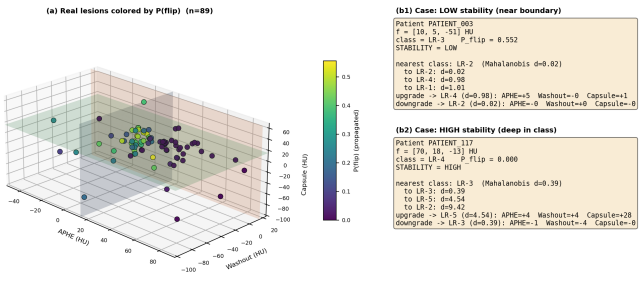

Figure 4: Decision panel on the 89 real lesions: (a) feature scatter coloured by P_{flip}; (b) case cards for an unstable LR-3 ($P_{\text{flip}} \approx 0.55$) and a confident LR-4 ($P_{\text{flip}} \approx 0$).

Table 3: Refinement effect on flip probability ΔP_{flip} by baseline stability tier (n=88); LOW row shows stabilized counts in parentheses.

tier (n)	α=0.5	α=1.0	α=1.5	α=2.0
LOW (5)	−0.017 (4/5)	−0.055 (5/5)	−0.074 (5/5)	−0.088 (5/5)
MEDIUM (42)	−0.000	+0.014	+0.013	+0.017
HIGH (41)	+0.009	+0.016	+0.022	+0.023
all (88)	+0.003	+0.011	+0.012	+0.014

A.3 Tiered Refinement Breakdown

A.4 Synthetic Perturbation Robustness (I3)

Under synthetic $\pm\{2, 5, 10\}$ HU feature noise, hard-threshold classification flips 51.7%/68.5%/89.9% of lesions ever, whereas the soft membership scores stay almost constant (Jensen–Shannon divergence 0.009/0.046/0.118).

Table 4: Synthetic feature-noise robustness on the 89-lesion cohort (I3).

perturbation	ever-flip rate	mean JS (soft)
±2 HU	51.7%	0.009
±5 HU	68.5%	0.046
±10 HU	89.9%	0.118