



ARTICLE

# Post-Hoc Texture–Ranking Fusion for Data-Efficient Hepatocellular Carcinoma Detection in Ultrasound Images

Hosoo Shin<sup>1</sup>, Seongyeon Son<sup>2</sup>, Minhee Park<sup>2</sup>, Lin Xia<sup>2</sup>, Yonghyun Lee<sup>3,\*</sup> and Eunchan Kim<sup>1,2,\*</sup>

<sup>1</sup>Department of Artificial Intelligence, Hanyang University, Seoul, Republic of Korea

<sup>2</sup>Department of Information Systems, Hanyang University, Seoul, Republic of Korea

<sup>3</sup>Department of Computer Engineering, Keimyung University, Daegu, Republic of Korea

\*Corresponding Authors: Yonghyun Lee. Email: yhlee@kmu.ac.kr; Eunchan Kim. Email: eckim@hanyang.ac.kr

Received: 15 July 2026; Accepted: 23 September 2026; Published: 21 October 2026

**ABSTRACT:** Data-scarce ultrasound-based cancer diagnosis remains a major computational challenge, and hepatocellular carcinoma (HCC) detection is a critical case in point: early diagnosis substantially improves patient outcomes, yet annotated data are severely limited and the residual errors of strong modern backbones concentrate in a narrow, near-boundary region. This study proposes a *post-hoc*, backbone-frozen fusion framework that improves an already-trained lesion classifier without any re-training or fine-tuning. A hand-crafted texture descriptor is injected into the frozen model's predictions through a lightweight class-balanced pairwise ranking scorer, based on Bayesian Personalized Ranking (BPR), and through a confidence-gated fusion rule that perturbs the prediction only in the uncertain, near-boundary region where the backbone is least reliable. The design offers four practical advantages: it needs no backbone re-training, fine-tuning, or architectural modification, so it can be attached to an already-validated classifier; it is model-agnostic, operating purely on output probabilities and input images; it is leakage-safe by construction, since scorer fitting, probability calibration, and hyper-parameter selection use only the training and validation partitions; and it is inexpensive, being fitted on cached probabilities and precomputed descriptors. Under ten-seed repeated 5-fold cross-validation on two architecturally distinct backbones, the diffusion classifier DiffMIC-v2 and the convolutional neural network (CNN)–Transformer MedViT-V2, the method raises sensitivity by 0.0221 ( $p < 0.001$ ) and 0.0455 ( $p < 0.01$ ), area under the curve (AUC) by 0.0030 ( $p < 0.05$ ) and 0.0150 ( $p < 0.001$ ), and F1 by 0.0090 ( $p < 0.001$ ) and 0.0269 ( $p < 0.01$ ), respectively, against a specificity cost of 0.0210 and 0.0145; in the pooled confusion matrices false negatives fall from 808 to 712 and from 1104 to 906. Ablations indicate that the improvement is driven mainly by the texture representation, in particular its local binary pattern (LBP) features, together with confidence gating, while the ranking scorer stabilizes the operating point and helps preserve specificity on the stronger backbone. This shift toward higher sensitivity could be advantageous for HCC screening, pending clinical validation.

**KEYWORDS:** Hepatocellular carcinoma detection; post-hoc fusion; Bayesian personalized ranking; texture descriptor; confidence-gated fusion

## 1 Introduction

Hepatocellular carcinoma (HCC) is the most common liver cancer, accounting for approximately 90% of all liver cancer cases, and it ranks as the third leading cause of cancer-related mortality worldwide [1,2]. When HCC is detected at an early stage, patient prognosis can be substantially improved and the 5-year survival rate may surpass 70%; however, detection at a progressed stage typically results in an average survival of only approximately one year. Stage at diagnosis is therefore the dominant determinant of outcome, and it is

the reason surveillance of at-risk populations is recommended: a large meta-analysis linked such surveillance to significantly higher early-stage detection, greater receipt of curative treatment, and longer overall survival [2,3]. The same logic drives early-detection research well beyond imaging, where photonic biosensing has been proposed as a complementary, label-free route: metamaterial and nano-photonic perfect-absorber biosensors have recently been designed to discriminate cancer-derived from normal circulating exosomes through refractive-index and terahertz-to-near-infrared spectral shifts [4,5]. These molecular and photonic modalities and image-based screening are complementary rather than competing: biomarker assays can flag who should be examined, whereas imaging must still localize and characterize the lesion. Improving the sensitivity of the imaging step, which is the objective of this work, therefore remains essential regardless of progress in biomarker-based detection.

Ultrasonography (US) is the most commonly used first-line imaging modality for HCC surveillance in high-risk patients. Ultrasonography offers several advantages over computed tomography (CT) and magnetic resonance imaging (MRI): relative to CT, it avoids ionizing radiation exposure, and relative to both, it offers higher accessibility and lower cost [6]. However, its sensitivity for identifying early-stage HCC is limited to 47%–63% [7,8], and in a prospective surveillance cohort it was as low as 27.3% for very early-stage tumors, defined as a single nodule smaller than 2 cm [9]. There are additional issues in that the diagnostic accuracy of ultrasonography depends largely on the clinician's experience and expertise, and image interpretation is often subjective, which may result in a non-negligible gap between observers. It is also well established that the diagnosis is largely influenced by patient factors, including anatomical blind spots within the liver, obesity, and morphological changes due to cirrhosis [10].

A fundamental source of this diagnostic difficulty lies in the low inter-class discriminability of ultrasound images at the pixel level. As illustrated in Fig. 1, benign liver lesions and malignant (HCC) lesions are visually indistinguishable across a wide range of pixel intensity levels. Statistical analysis of the dataset used in this study shows that 99.5% of malignant images fall within the pixel intensity range of benign images: the malignant range is almost entirely contained in the benign range, even though the difference in means is itself non-trivial (Cohen's  $d = 0.69$ ). Raw intensity therefore separates the two classes poorly, since a given intensity value is largely uninformative about the label. This low signal-to-noise ratio at the feature level explains why pixel statistics alone are insufficient for reliable classification, and motivates the need for representation learning methods capable of capturing higher-order discriminative structure in the latent space.

To address this limitation, several recent approaches have attempted to improve the accuracy and consistency of HCC diagnosis by implementing deep learning methods for ultrasonography [11–13]. Specifically, artificial intelligence (AI)-based research on ultrasonography is evolving toward a hybrid model approach that integrates multiple architectures with a staged pipeline, which is progressing toward enhancing clinical applicability and improving model performance [14].

For instance, She et al. [15] proposed the HSQformer, which achieved senior radiologist-level accuracy in early HCC detection, effectively integrating a CNN module for local feature extraction and a Transformer module for global context extraction. Lu et al. [16] improved the diagnostic specificity and clinical work efficiency of physicians by building a multi-stage pipeline that fused UniMatch and LivNet, which were utilized for lesion detection and diagnostic classification, respectively. Rhyou and Yoo [17] improved model performance by successively applying preprocessing methods tailored to the characteristics of ultrasound images and an object detection algorithm.

Although the efficacy and significance of AI-enhanced ultrasound lesion diagnosis have been highlighted, data sparsity remains the main impediment to developing a precise AI model for diagnosis [18,19]. This is primarily because medical image data acquisition is costly and time-intensive, and the labeling process requires the supervision of medical professionals.

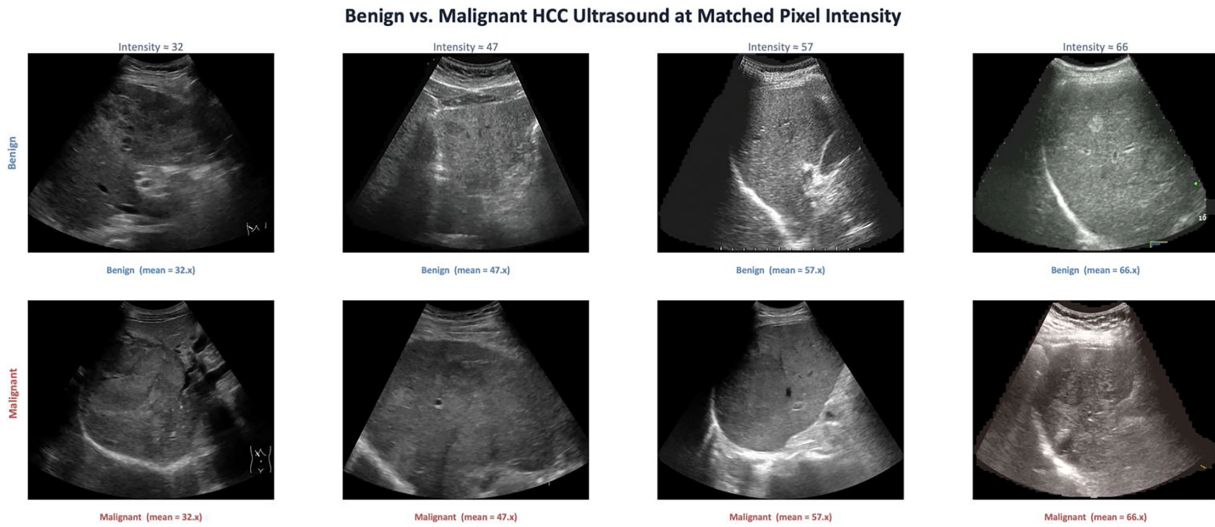


**Figure 1:** Liver ultrasound images of benign lesions and malignant (HCC) lesions matched by mean pixel intensity (columns: intensity  $\approx$  32, 47, 57, 66). Despite belonging to different classes, the images are visually indistinguishable across intensity levels, illustrating the fundamental challenge of pixel-level discrimination. Statistical analysis confirms that 99.5% of malignant samples fall within the pixel intensity range of benign samples (Cohen's  $d = 0.69$ ).

Privacy and legal concerns further complicate data acquisition. Medical datasets that include sensitive patient health information are protected by several strict regulations, including the Health Insurance Portability and Accountability Act (HIPAA). Many patients are reluctant to disclose their medical information due to concerns regarding confidentiality and integrity [20,21].

Faced with the abovementioned issues, access to medical datasets remains a major hurdle for deep learning researchers [18,22]. This is particularly prominent in the field of ultrasound imaging because of the lack of standardization of the examination process and the fundamental physical constraints of ultrasound imaging [23]. Thus, the construction of a consistent and standardized dataset is more challenging.

To overcome these limitations in existing research, we focused on training under a more realistic and challenging environment with a small HCC ultrasound image dataset, and chose MedViT-V2 [24] and DiffMIC-v2 [25] as our primary backbones, representative state-of-the-art medical image classification models with architecturally distinct designs. A central empirical observation motivates our approach: in our experiments the residual errors of these strong backbones concentrate in a narrow, near-boundary region where the raw pixel distributions of the two classes overlap almost completely (Fig. 1), rather than being spread across the feature space. Because the errors are localized in this way, an intervention that acts selectively in that region can address a large share of them. Modifying the training objective (for example, with ranking-, contrastive-, or triplet-style representation-learning losses applied inside the network) is a conceivable alternative route, but it requires re-training and re-validating the backbone and introduces further design choices; we do not evaluate such training-time objectives in this work, and we make no claim about how they would compare with the post-hoc approach studied here.

We therefore shift the intervention from training time to *after* training. We propose a post-hoc, backbone-frozen fusion framework: the backbone is trained once and frozen, and a complementary, hand-crafted texture representation—a partially independent view of local echo texture—is injected into its predictions. The injection uses a lightweight class-balanced pairwise ranking scorer, based on Bayesian Personalized Ranking (BPR), that orders malignant above benign samples in texture space and is fit only on the training partition, followed by a confidence-gated fusion rule that alters the prediction only where the

backbone is uncertain, leaving confident predictions intact. Because scorer fitting, calibration, and hyperparameter selection use only training and validation data, the procedure is leakage-safe; because it operates purely on output probabilities and input images, it is model-agnostic and costs a fraction of a re-training run.

Using the proposed method, we provide a lightweight framework that can recover some of the subtle, near-boundary information a frozen backbone may miss under scarce-data conditions in this noise-prone domain.

This study makes the following key contributions:

- We characterize the error structure of strong backbones on small ultrasound HCC data, showing quantitatively that their residual errors concentrate in a narrow near-boundary band covering roughly 34%–41% of cases, and that a complementary hand-crafted texture view is competitive with, and on the weaker backbone superior to, the backbone precisely inside that band. This localization of the errors, together with the practical requirement that an already-trained and validated classifier should not have to be re-trained, motivates a post-hoc intervention.
- We propose a post-hoc, backbone-frozen fusion framework whose core is a complementary hand-crafted texture representation combined through a confidence-gated fusion rule that modifies predictions only in the uncertain region so that confident decisions are preserved; the texture signal and the gating are the primary drivers of the improvement (within the texture signal, local binary pattern features are the dominant family), and the framework is fit without test-set leakage. As the scorer that maps texture to a probability, we adopt a class-balanced BPR ranking objective which, unlike a non-parametric prototype alternative, controls the operating-point shift and preserves specificity and balanced accuracy on already-strong backbones. In contrast to prior BPR applications, which inject the ranking objective during training on graph-structured biological entities, ours applies BPR entirely post-hoc on image data as a complement to a frozen backbone.
- We validate the framework on two architecturally distinct backbones, the diffusion classifier DiffMIC-v2 and the CNN–Transformer MedViT-V2, under ten-seed repeated 5-fold cross-validation, and observe statistically significant and consistent gains in sensitivity, AUC, and F1, suggesting that the benefit is not tied to a single architecture, at least across the two backbones evaluated here.

## 2 Related Work

### 2.1 Artificial Intelligence and Image Processing in the Diagnosis, Treatment, and Prognosis of Liver Cancer

Artificial intelligence (AI) now spans the entire clinical trajectory of liver cancer, and it is useful to separate the three stages the present work touches on, since our contribution belongs squarely to the first.

For *diagnosis*, deep learning has been applied most heavily to lesion detection and characterization from imaging: a detection-transformer trained on abdominal ultrasound achieved high sensitivity and specificity for lesion detection while performing on par with expert readers at benign-vs.-malignant characterization [26]. Machine-learning and radiomic models have likewise been applied to *treatment*-response prediction and to *prognosis*. A comprehensive review of AI across HCC prevention, surveillance, diagnosis, treatment selection, and prognostication situates these strands and identifies data standardization, model generalizability, and interpretability as the principal obstacles to clinical adoption [27].

Classical image processing has not been displaced by this progress; it has been absorbed into it. Quantitative texture analysis (first-order histogram statistics, gray-level co-occurrence matrices, run-length matrices, and related descriptors) remains a mainstay of HCC characterization on CT and MRI for diagnosis, grading, treatment response, and prognosis, with reproducibility and acquisition variability as the recognized limiting factors [28]. The two families are complementary in a specific sense that matters for our design:

a learned representation is optimized end-to-end for the training distribution and can therefore inherit that distribution's biases, whereas a deterministic texture descriptor is fixed in advance, carries no trainable parameters, and consequently provides a view that is only partially correlated with the backbone's. Our method exploits this property. Rather than replacing the deep representation with hand-crafted features, or concatenating the two before a jointly trained classifier, we keep the trained classifier untouched and inject the classical descriptor only where the deep model's own evidence is weakest, thereby placing our method at the diagnosis stage of the pipeline above while remaining deployable on top of any of the diagnostic models it surveys.

## 2.2 Metric Learning in Medical Imaging

**Distance-Based Metric Learning Methods and Their Medical Applications.** Metric learning is a training strategy that brings semantically similar sample pairs close together, and separates dissimilar pairs in the embedding space. Medical imaging has widely adopted this approach to enhance class separability and improve generalization performance even in data-scarce environments [29].

Contrastive loss [30] focuses on ensuring the distance between classes by minimizing the distance between positive pairs and maximizing the distance between negative pairs. Triplet loss [31], which directly learns the relative distance between an anchor and both its positive and negative samples, improves performance by distinguishing fine lesion boundaries. Later, center loss [32] improved the intra-class compactness by learning the centroid for each class, and Proxy-NCA [33] enhanced the computational performance by employing a proxy vector to represent each class.

This distance-based learning demonstrated a high generalization performance in pathological slide retrieval, disease severity classification, and similarity-based diagnostic tasks. It also holds clinical significance because it enables representation learning that reflects pathological similarities beyond mere label classification [29,32].

## 2.3 Ranking-Based Approaches in Medical Representation Learning

While metric learning learns the correlation between samples based on the absolute distance or margin, ranking-based learning directly models the relative ranking between samples. This approach encourages the model to boost the ranking of semantically meaningful sample pairs, while lowering the ranking of less meaningful sample pairs. For instance, ConPro [34] demonstrated that by learning the ranking of lesion severity between images, a model can detect more subtle clinical distinctions than an approach that simply distinguishes between class labels.

Mazaheri et al. [35] achieved the mitigation of institutional bias by applying ranking loss for pathology image retrieval and further demonstrated a performance improvement of approximately 5%. These results demonstrate that ranking-based learning is well-suited for medical imaging environments in which class boundaries are ambiguous or label noise exists. Thus, the pairwise ranking loss, which provides more flexible decision boundaries for medical imaging representation learning, has been highlighted as an effective alternative for alleviating data imbalance and domain bias.

## 2.4 Bayesian Personalized Ranking in Biomedical Applications

The Bayesian Personalized Ranking (BPR) loss [36] is a loss function originally designed to learn candidate item rankings based on user preferences in recommender systems. This loss encourages the model to predict that the score of the preferred item  $i$  is higher than that of the unpreferred item  $j$  ( $r_i > r_j$ ) for every positive-negative pair, which allows the model to learn the ranking. BPR loss is formulated as follows:

$$\mathcal{L}_{bpr} = -\frac{1}{|S|} \sum_{(i,j) \in S} \log \sigma(r_i - r_j), \quad (1)$$

where  $S$  is the set of positive-negative pairs  $(i, j)$  formed within a batch,  $r_i$  and  $r_j$  represent the predicted scores for the preferred sample  $i$  and the unpreferred sample  $j$ , respectively, and  $\sigma(\cdot)$  denotes the sigmoid function. This loss function aims to maximize the probability of positive pairs being ranked higher than negative pairs.

Peska et al. [37] applied BPR in recommender systems to drug-target interaction (DTI) prediction to rank potential target proteins for each drug. This is a representative case in which BPR has been practically utilized for medical and bioinformatics data analysis, which shows the potential of pairwise ranking-based learning in the medical domain.

More recent work has extended BPR to a broader range of biomedical prediction tasks framed as ranking problems over biological association graphs. CLTDA (contrastive learning for transfer-RNA-derived small RNA (tsRNA)-disease association) [38] combines BPR loss with contrastive learning on graph convolutional networks to predict tsRNA-disease associations, employing Bayesian negative sampling to reduce false negatives during training. AGDNGDA (adaptive graph diffusion network for gene-drug association) [39] uses BPR as the training objective for an adaptive graph diffusion network that models gene-drug associations via heat-kernel-based neighborhood aggregation. DriverSub-SVM [40] applies BPR to infer personalized driver gene rankings for individual cancer patients from multi-omics data, then aggregates these into consensus features for subtype classification using a support vector machine (SVM). Concurrently, a graph-theoretic analysis [41] has demonstrated a fundamental limitation of standard BPR: the inner-product score is equivalent to counting 2-hop paths in an embedding network, capturing only local connectivity while missing global topological structure.

Despite their diversity, all existing biomedical applications of BPR share a common characteristic: they operate on graph-structured biological entities (genes, drugs, RNAs, diseases), using BPR as an entity-level ranking signal, and they inject the ranking objective *during* training. None address medical image classification in a post-hoc, backbone-frozen setting. In this work, we repurpose the pairwise ranking discipline of BPR as a lightweight, post-hoc scorer over a complementary texture representation: rather than reshaping a backbone's internal features during training, we fit a class-balanced BPR scorer that orders malignant above benign samples in texture space and fuse its calibrated probability into a frozen backbone's prediction only where the backbone is uncertain. This use of BPR as an external, texture-based ranking signal is well suited to data-scarce medical imaging tasks where strong backbones are already competitive and their residual errors tend to concentrate near the decision boundary. The approach also relates to, but does not subsume, two established directions. Ranking- and preference-based learning has been used in medical imaging, but to shape representations during training or to reduce retrieval bias rather than to fuse predictions post-hoc [34,35]. Likewise, hand-crafted texture and radiomic descriptors are frequently combined with deep representations, but typically through feature-level fusion or ensembling rather than by injecting a calibrated texture probability only where the backbone is uncertain. Methodologically, the closest relatives are post-hoc calibration methods such as temperature and Platt scaling [42], which likewise adjust a frozen model's outputs without re-training; unlike those methods, which only rescale the existing decision function, ours introduces a complementary representation through a ranking objective. To our knowledge, this is the first use of a BPR-style pairwise ranking scorer as a post-hoc, texture-based fusion signal for frozen medical image classifiers.

## 2.5 Texture Descriptors and Multiresolution Texture Analysis

The texture descriptor used in this work is deliberately conventional: first-order intensity statistics, second-order Haralick features computed from the gray-level co-occurrence matrix (GLCM) [43], and a rotation-invariant local binary pattern (LBP) histogram [44]. Richer descriptor families exist and have a substantial track record in medical image analysis. Multiresolution descriptors derived from Gabor filter banks summarize the energy of an image across a set of scales and orientations [45]. Wavelet-based descriptors have been applied directly to the problem studied here: wavelet-packet sub-band texture features combined with a support vector machine (SVM) classified normal liver, cirrhosis, and HCC from B-mode ultrasound regions of interest with high overall accuracy [46]. A second direction generalizes the co-occurrence idea itself: generalized or superior-order co-occurrence matrices extend the classical second-order GLCM by jointly encoding additional image attributes or by considering more than two pixels at a time. Most directly related to the present study, this descriptor family has been combined with convolutional networks for HCC recognition from ultrasound, with textural and CNN-derived features fused at the classifier level [47].

Our aim in this work is not to identify the strongest possible texture representation. It is to test whether a compact, deterministic, training-free descriptor, one that adds no trainable parameters in a severely data-limited setting, carries enough complementary evidence to correct a frozen backbone near its decision boundary. The 57-dimensional descriptor is therefore chosen for reproducibility and low capacity rather than expressiveness, and the texture-component ablation of [Section 4.7](#) quantifies what each family contributes. Because the fusion machinery is agnostic to how  $\phi(x)$  is computed, substituting or augmenting it with Gabor, wavelet-packet, or generalized co-occurrence features is a direct extension, and we return to this point in [Section 5.6](#).

## 2.6 Adjacent Research Directions in Liver Disease and Liver Cancer

Beyond the imaging and representation-learning literature that our method builds on, several adjacent research directions shape the clinical environment in which an automated HCC screening tool would operate. We summarize them briefly here; they are context for the problem rather than components of the proposed method.

**The Internet of Things in medical practice.** Technological development in the medical field has been rapid and continues to evolve, and one of the most consequential recent shifts has been the introduction of the Internet of Things (IoT) into medical practice. In hepatology, IoT-linked frameworks have begun to couple continuous physiological monitoring with explainable classifiers for liver-disease diagnosis [48]. For liver cancer specifically, the relevance is largely infrastructural: IoT platforms determine how surveillance data are captured and moved, and a decision-support layer such as ours, which requires only a trained model's output probabilities and the image itself, is comparatively easy to place at the edge of such a pipeline.

**Hepatocolic fistula after ablative radiation lobectomy.** Locoregional therapy of hepatic malignancy carries rare but serious delayed complications. Hepatocolic fistula has been reported when ablative radiation lobectomy was used to treat a right-lobe colorectal metastasis [49]. Such complications underline that imaging of the liver does not end at diagnosis: the same modality is relied upon for post-treatment surveillance, where the appearances are unfamiliar and the diagnostic question differs from the one considered here.

**Toll-like receptors in liver regeneration after partial hepatectomy.** Resection remains a curative option for HCC, and its feasibility depends on the regenerative capacity of the remnant liver. Toll-like receptor (TLR) signaling is a central mediator of that process: hepatic TLR expression rises significantly in the hours following partial hepatectomy [50]. This biology bears on the value of early detection from the

opposite direction to imaging: the earlier a lesion is identified, the more likely resection is to be technically feasible and the regenerative reserve sufficient.

**Primary squamous cell carcinoma of the liver.** This tumor is a rare primary hepatic malignancy with a poor prognosis, typically reported as mean survival under one year, and a histogenesis that remains debated [51]. It is of direct methodological interest here because of its imaging behavior. A multimodality review found that conventional ultrasound shows only a mixed-echogenicity mass and that CT, MRI, and positron emission tomography–computed tomography (PET-CT) appearances are similarly non-specific, so the lesion is routinely mistaken for HCC, cholangiocarcinoma, or a liver abscess, with definitive diagnosis requiring histopathological confirmation [52]. This is the same failure mode that motivates our work: when lesions of different pathology occupy overlapping regions of the intensity distribution, decisions based on gross appearance are unreliable, and finer descriptions of internal echo structure become the most discriminative signal available from imaging itself, even if definitive diagnosis ultimately rests on histopathology.

**Three-dimensional printing in liver cancer.** 3D-printed liver models, reconstructed from a patient’s own segmented CT or MRI, have been used for preoperative planning, surgical simulation, and training in hepatic surgery. One report is particularly pertinent to the limits of ultrasound: 3D-printed models were used to guide hepatectomy for small metastases that had become invisible on intraoperative ultrasonography after chemotherapy [53]. Both 3D printing and the present work respond to the same underlying difficulty, namely that small or low-contrast hepatic lesions are hard to localize reliably from ultrasound alone, but they do so at different points, one by supplying an alternative spatial representation in the operating room and the other by extracting more evidence from the ultrasound image itself.

### 3 Materials and Methods

We propose a *post-hoc*, backbone-frozen fusion framework that improves an already-trained lesion classifier without any modification, re-training, or fine-tuning of the backbone. [Section 3.1](#) gives an overview and motivation; [Section 3.2](#) fixes the problem setting and notation; [Section 3.3](#) defines the complementary texture representation; [Section 3.4](#) introduces the class-balanced BPR scorer; [Section 3.5](#) presents the confidence-gated fusion rule; and [Section 3.6](#) describes leakage-safe hyper-parameter selection. [Fig. 2](#) gives an overview of the full pipeline. Additional details on the texture descriptor, hyper-parameter selection, and paired evaluation protocol are provided in [Appendix A](#).

#### 3.1 Overview and Motivation

Two considerations motivate the design. The first is empirical: on small, noisy ultrasound HCC data the residual errors of the modern backbones we study concentrate in a narrow, near-boundary region rather than being spread across the feature space (quantified in [Section 4.10](#)), so an intervention that acts selectively in that region can address most of the remaining errors. The second is practical: any intervention on the training objective requires re-training and re-validating the backbone, whereas a post-hoc layer can be attached to an already-trained, already-validated model at negligible cost and removed just as easily. Rather than competing with the backbone on the axis it already models well, we therefore introduce a hand-crafted texture axis *after* training and apply it only where the backbone is uncertain.

[Fig. 2](#) summarizes how these parts fit together. The diagram is a hand-drawn schematic of the implemented pipeline rather than an automatically generated computation graph: each block was placed to correspond one-to-one with a stage of the procedure, and the arrows indicate the direction of data flow at inference time. The essential structural feature is that the two paths leaving the input image are independent: the backbone never sees the texture descriptor and the texture scorer never sees the backbone’s features,

so the only point of interaction is the fusion block, and the only quantity that controls that interaction is the backbone's own confidence. This is what makes the framework model-agnostic: replacing the backbone changes the upper path alone, and everything downstream continues to operate on a scalar probability.

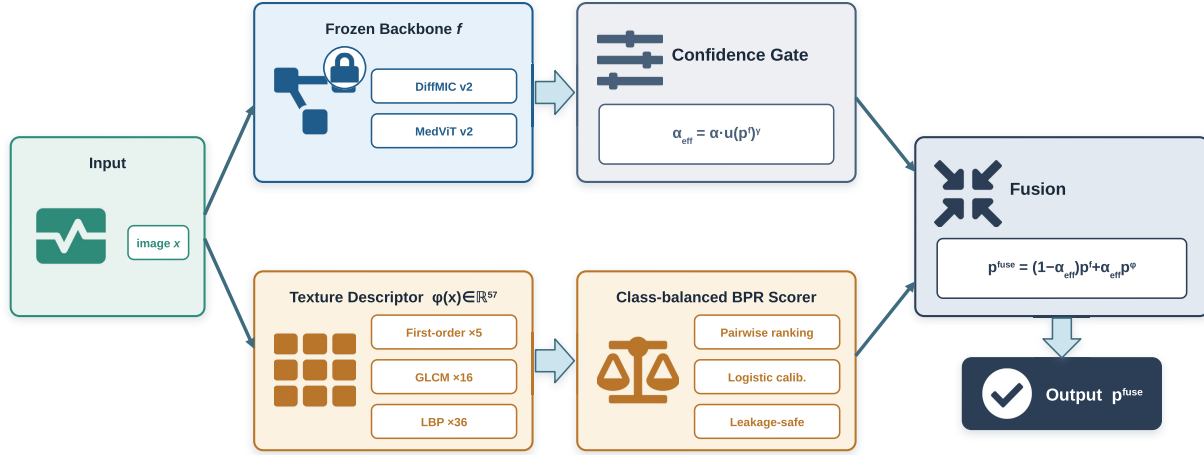


**Figure 2:** Overview of the proposed post-hoc texture–ranking fusion. The diagram is a schematic drawn by the authors to mirror the implemented pipeline block by block; it is not the output of an automatic model-visualization tool, and every block corresponds to a component defined in Sections 3.3–3.6. Reading left to right, a single input image  $x$  is passed along two independent paths. On the upper path the frozen backbone  $f$  (DiffMIC-v2 or MedViT-V2, weights locked) produces the probability  $p^f$ , which is also fed to the confidence gate that computes  $\alpha_{\text{eff}} = \alpha u(p^f)^\gamma$ . On the lower path the same image yields the 57-dimensional texture descriptor  $\phi(x)$  (5 first-order + 16 GLCM + 36 LBP components), which the class-balanced BPR scorer maps to a scalar score and a logistic calibration converts to  $p^\phi$ ; both are fit on the training partition only. The two paths meet at the fusion block, which returns  $p^{\text{fuse}} = (1 - \alpha_{\text{eff}})p^f + \alpha_{\text{eff}}p^\phi$ . Because  $\alpha_{\text{eff}}$  vanishes when  $p^f$  is close to 0 or 1, the lower path can only influence the output where the upper path is uncertain.

### 3.2 Problem Setting and Notation

Let  $\mathcal{D} = \{(x_i, y_i)\}$  be a binary lesion dataset with  $y_i \in \{0, 1\}$  (0: benign, 1: malignant). Within each cross-validation fold the data are split into training, validation, and test subsets  $\mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{val}}, \mathcal{D}_{\text{te}}$ ; the backbone is trained on  $\mathcal{D}_{\text{tr}}$  with early stopping on  $\mathcal{D}_{\text{val}}$  and is never modified thereafter. The post-hoc scorer, its probability calibration, and the fusion hyper-parameters are all estimated on  $\mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{val}}$ ;  $\mathcal{D}_{\text{te}}$  is used exclusively for final evaluation.

### 3.3 Complementary Texture Representation

HCC discrimination on ultrasound depends heavily on internal echo texture—speckle, local heterogeneity, and boundary echo patterns—not on absolute intensity alone. We capture this with a hand-crafted texture representation that offers a partially independent view, one that may not be fully reflected in the backbone probability. We deliberately use a compact, deterministic descriptor instead of an additional learned module: the aim is to add a low-capacity auxiliary view under scarce-data conditions without introducing another trainable image encoder. For each image we compute a training-free descriptor  $\phi(x) \in \mathbb{R}^{57}$  consisting of three groups:

- **First-order statistics (5):** mean, standard deviation, skewness, kurtosis, and entropy of the pixel-intensity histogram.

- **Gray-level co-occurrence matrix, GLCM (16):** contrast, homogeneity, energy, and correlation, each computed at  $L = 16$  quantized gray levels over the four principal directions at unit distance  $\{(0, 1), (1, 0), (1, 1), (1, -1)\}$ .
- **Local binary patterns (36):** a 36-bin rotation-invariant LBP histogram.

These groups are drawn from long-established texture families—first-order intensity statistics, Haralick gray-level co-occurrence features [43], and rotation-invariant local binary patterns [44], all standard descriptors in image and radiomics analysis [54]. The 57-dimensional vector is thus a compact, conventional texture descriptor rather than an exhaustive radiomics feature set; we deliberately keep it low-capacity to limit overfitting in the small-data setting. The descriptor is computed identically for either backbone, which makes the downstream fusion model-agnostic. It is not meant to replace the backbone representation, but to supply a low-dimensional secondary view whose complementarity is tested during post-hoc fusion.

### 3.4 Class-Balanced BPR Ranking Scorer

On the training fold we standardize the descriptor to  $\tilde{\phi}$  (mean and variance from  $\mathcal{D}_{\text{tr}}$  only) and learn a scalar score  $s(\tilde{\phi}) = w^\top \tilde{\phi} + b$ . Our primary scorer is formulated as a class-balanced, asymmetric-margin *pairwise ranking* objective in the style of BPR [36]. We emphasize that Eq. (2) is not a new derivation: it is the standard BPR pairwise criterion, in which the log-sigmoid is written equivalently as a softplus, with two established modifications, an asymmetric margin  $m$  and the effective-number class weighting of Cui et al. [55], assembled for the present setting. It reads

$$\mathcal{L}_{\text{BPR}} = \mathbb{E}_{(p,n) \sim \mathcal{P} \times \mathcal{N}} \left[ \omega \text{softplus}(- (s_p - s_n - m)) \right], \quad (2)$$

where  $\mathcal{P}, \mathcal{N}$  are the malignant/benign index sets,  $s_p, s_n$  the scores of a malignant–benign pair,  $m$  an asymmetric margin,  $\text{softplus}(z) = \log(1 + e^z)$  is the smooth approximation to the rectifier, and  $\omega$  is the effective-number class weight  $\omega_c \propto (1 - \beta)/(1 - \beta^{n_c})$  with  $\beta = 0.999$  of Cui et al. [55]. Setting  $m = 0$ ,  $\omega \equiv 1$ , and  $\text{softplus}(-z) = -\log \sigma(z)$  recovers Eq. (1) exactly, so Eq. (2) is a re-parameterization of the original BPR criterion rather than a separate loss. It optimizes the *relative ordering* of malignant over benign samples in texture space and is fit by stochastic gradient descent over sampled pairs. The raw score is then mapped to a probability  $p^\phi$  by a one-dimensional logistic calibration fit on  $\mathcal{D}_{\text{tr}}$ , following the standard Platt-scaling procedure for turning uncalibrated scores into probabilities [56].

Retaining the pairwise ranking formulation of the original framework, we additionally evaluate two prototype-aware scorer variants as an ablation: a non-parametric *Proto* score, the signed cosine proximity of  $\tilde{\phi}$  to the malignant vs. benign class prototype estimated by the geometric median, a robust multivariate location estimator computed by Weiszfeld’s iteration in the modified form of Vardi and Zhang [57], and a *Hybrid* score that averages the standardized BPR and Proto scores. The BPR scorer is the default; the prototype variants isolate how much of the effect is due to the ranking objective itself.

### 3.5 Confidence-Gated Fusion

Uniformly averaging  $p^f$  and  $p^\phi$  corrupts confident, correct predictions and collapses specificity, because the texture signal is weaker than the backbone on easy cases. We therefore fuse only where the backbone is uncertain. With an uncertainty gate  $u(p)$  that is maximal at the boundary  $p = 0.5$  and vanishes at the extremes—triangular  $u(p) = 1 - 2|p - 0.5|$  or binary entropy—the fused probability is

$$\alpha_{\text{eff}} = \alpha u(p^f)^\gamma, \quad p^{\text{fuse}} = (1 - \alpha_{\text{eff}}) p^f + \alpha_{\text{eff}} p^\phi. \quad (3)$$

Confident predictions ( $p^f \rightarrow 0, 1$ ) satisfy  $\alpha_{\text{eff}} \rightarrow 0$  and are preserved; ambiguous, near-boundary predictions receive the complementary texture signal in proportion to the backbone's uncertainty. The exponent  $\gamma$  sharpens the gate so that only the most uncertain cases are perturbed. Fig. 3 plots the resulting weight. The curves are obtained analytically rather than experimentally: Eq. (3) was evaluated in closed form on a dense grid of  $p^f \in [0, 1]$  with  $\alpha$  fixed at 0.3, for the triangular gate at  $\gamma \in \{2, 3, 4\}$  and for the entropy gate at  $\gamma = 3$ , and the resulting values of  $\alpha_{\text{eff}}$  were plotted directly. No model, data, or fitting is involved, so the figure characterizes the mechanism itself and is identical for both backbones.

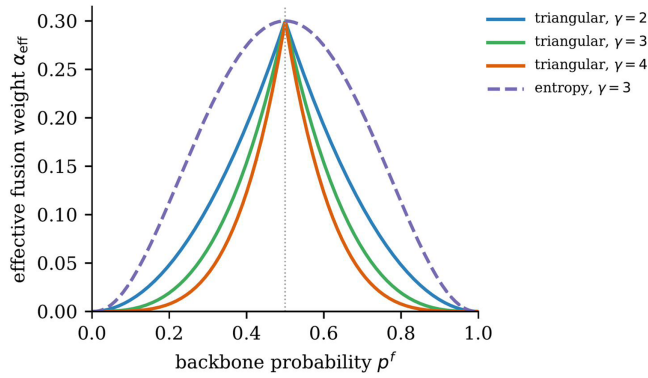


**Figure 3:** Effective fusion weight  $\alpha_{\text{eff}} = \alpha u(p^f)^\gamma$  vs. backbone probability  $p^f$ , for the triangular gate at  $\gamma \in \{2, 3, 4\}$  and the entropy gate at  $\gamma = 3$ , all at  $\alpha = 0.3$ . The curves are the closed-form evaluation of Eq. (3) on a dense grid of  $p^f \in [0, 1]$  and involve no data or model fitting. All curves peak at  $\alpha_{\text{eff}} = \alpha$  at the decision boundary  $p^f = 0.5$  (dotted vertical line) and vanish at the confident extremes, so confident predictions are preserved by construction. Larger  $\gamma$  narrows the region of intervention without changing its maximum strength, whereas the entropy gate is broader than the triangular gate at the same  $\gamma$ .

Three properties are visible and each has a design consequence. First, every curve peaks at exactly  $\alpha_{\text{eff}} = \alpha$  at  $p^f = 0.5$  and decays to zero at both extremes, so the mixing weight is bounded above by  $\alpha$  everywhere: with the conservative  $\alpha = 0.2$  selected in roughly half of all folds (Table 1), the texture path can never contribute more than one fifth of the fused probability. Second, increasing  $\gamma$  from 2 to 4 leaves the peak untouched but contracts the region over which the weight is non-negligible; at  $\gamma = 4$ , the weight has already fallen to about a tenth of its maximum by  $p^f \approx 0.3$ , whereas at  $\gamma = 2$  it is still around a third. The exponent therefore behaves as a width control on the region of intervention and not as a strength control. Third, the entropy gate is visibly broader than any of the triangular gates, because binary entropy decays more slowly than  $1 - 2|p - 0.5|$  near the boundary; it consequently spreads the correction over more cases. Since neither gate dominates the other a priori, both are left in the search grid and the choice is made per fold on validation AUC, which is why both are selected in practice (Section 4.9).

**Table 1:** Frequency (out of 50 folds; ten seeds  $\times$  five folds) with which each fusion hyper-parameter value was selected on the validation partition, per backbone. Gate  $\in \{\text{triangular, entropy}\}$ , mixing weight  $\alpha$ , and gate exponent  $\gamma$ . Selection maximizes validation AUC and never consults test labels.

|      | Value      | DiffMIC-v2 | MedViT-V2 |
|------|------------|------------|-----------|
| Gate | Triangular | 34         | 33        |
|      | Entropy    | 16         | 17        |

(Continued)

**Table 1 (continued)**

|                        | Value | DiffMIC-v2 | MedViT-V2 |
|------------------------|-------|------------|-----------|
| Mixing weight $\alpha$ | 0.2   | 24         | 26        |
|                        | 0.3   | 11         | 7         |
|                        | 0.5   | 15         | 17        |
| Gate exponent $\gamma$ | 2     | 14         | 19        |
|                        | 3     | 5          | 13        |
|                        | 4     | 31         | 18        |

### 3.6 Leakage-Safe Hyper-Parameter Selection

The fusion hyper-parameters  $(\alpha, \gamma, \text{gate})$  are selected automatically *per fold* on the validation partition: with the scorer fixed, we choose the combination of  $\alpha \in \{0.2, 0.3, 0.5\}$ ,  $\gamma \in \{2, 3, 4\}$ , and  $\text{gate} \in \{\text{triangular}, \text{entropy}\}$  that maximizes validation AUC, and apply it once to the test partition. Scorer weights, logistic calibration, and fusion hyper-parameters are thus all determined without consulting test labels, so the test partition constitutes a held-out evaluation. Algorithm 1 summarizes the procedure.

---

**Algorithm 1:** Post-hoc texture–ranking fusion

---

**Input:** Frozen backbone  $f$ ; training set  $\mathcal{D}_{\text{tr}}$ ; validation set  $\mathcal{D}_{\text{val}}$ ; texture map  $\phi$

**Output:** Fused probability  $p^{\text{fuse}}$

- 1: **Texture Feature Extraction**
  - 2: Compute  $\phi(x)$  for all images; standardize using  $\mathcal{D}_{\text{tr}}$  statistics
  - 3: **Scorer Fitting**
  - 4: Fit class-balanced BPR scorer  $s(\cdot)$  on  $\mathcal{D}_{\text{tr}}$  (Eq. (2))
  - 5: Fit 1-D logistic calibration  $s(\cdot) \rightarrow p^\phi$  on  $\mathcal{D}_{\text{tr}}$
  - 6: **Hyper-Parameter Selection**
  - 7: Select  $(\alpha^*, \gamma^*, \text{gate}^*)$  maximizing AUC on  $\mathcal{D}_{\text{val}}$  (Eq. (3))
  - 8: **Confidence-Gated Fusion**
  - 9: Set  $\alpha_{\text{eff}} = \alpha^* u(p^f)^{\gamma^*}$  with the selected gate
  - 10: **return**  $p^{\text{fuse}} = (1 - \alpha_{\text{eff}}) p^f + \alpha_{\text{eff}} p^\phi$
- 

## 4 Experiments and Results

### 4.1 Dataset

All experiments use a publicly available liver ultrasound image dataset provided by the Third Affiliated Hospital of Sun Yat-sen University [15], comprising 635 lesion images labeled as benign (200) or malignant (435). The same dataset and the same partitions are used for every backbone and for its post-hoc counterpart, ensuring a strictly paired comparison. This dataset was introduced in an arXiv preprint [15] that also released an associated model implementation<sup>1</sup>. As the authors state that the paper has not been formally published and the code has not been organized, with updates pending, we treat this work as dataset provenance and related context rather than a primary benchmark, and base our comparisons on the formally published, reproducible backbones DiffMIC-v2 [25] and MedViT-V2 [24].

<sup>1</sup><https://github.com/Asunatan/HSQformer>

## 4.2 Evaluation Protocol

**Data augmentation.** Because the dataset is small (635 images), we state the augmentation protocol explicitly. Augmentation is applied *on the fly during backbone training only*, to the training partition of each fold; it is never applied to the validation or test partitions, and no offline or synthetic augmentation is performed, so the number of distinct images is 635 throughout and no augmented copy of a training image can appear in a test partition. Each backbone is trained with the default augmentation pipeline of its original implementation (standard geometric and photometric transforms applied to the input resized to  $224 \times 224$ ), with the exact settings given in the released code. Both backbones use ImageNet channel normalization. Validation and test images undergo only deterministic preprocessing (resizing and normalization), with no random augmentation. The texture descriptor  $\phi(x)$  is likewise computed on the deterministic, non-augmented image, so the descriptor for a given image is fixed across seeds and folds and the post-hoc stage introduces no additional stochasticity. Since augmentation acts only on the frozen backbone's training and the post-hoc layer is fitted on cached probabilities and precomputed descriptors, the reported  $\Delta$  values isolate the effect of the fusion and not of the augmentation policy.

**Cross-validation.** We use repeated stratified cross-validation with ten independent repetitions. For each seed the data are independently re-partitioned into a 5-fold split, giving 50 train/validation/test partitions in total; re-randomizing the partition at every seed reduces the influence of any single favorable split. For each fold the backbone is trained and frozen, and the post-hoc scorer, calibration, and fusion hyper-parameters are estimated on the training and validation partitions only (Section 3.6). A backbone and its post-hoc counterpart share *identical* partitions at every seed and fold, so the two are compared in a paired fashion. We report per-metric mean  $\pm$  standard deviation (SD) across the ten seeds and assess significance with a paired  $t$ -test on the seed-level differences  $\Delta$  between the post-hoc method and its own baseline. We report accuracy (ACC), balanced accuracy (BACC), AUC, sensitivity (recall), specificity, precision, and F1.

## 4.3 Backbones

We evaluate the framework on two architecturally distinct, state-of-the-art medical image classification backbones, so that any consistent gain can be attributed to the model-agnostic post-hoc procedure rather than to a single architecture. We do not include HSQformer—the task-specific model released with the dataset (Section 4.1)—as a primary quantitative baseline, since adapting its actively developed preprint implementation to our exact repeated 5-fold protocol would introduce additional implementation choices that could confound a controlled comparison; we therefore discuss it as related work and dataset provenance.

**DiffMIC-v2.** DiffMIC-v2 [25] is a dual-guidance diffusion network for medical image classification, offering strong classification performance and enhanced training stability. It is the stronger of the two baselines on this dataset.

**MedViT-V2.** MedViT-V2 [24] is a CNN–Transformer hybrid backbone that integrates Kolmogorov–Arnold Networks with dilated neighborhood attention, replacing global self-attention with spatially adaptive local attention. It is the weaker of the two baselines on this dataset.

## 4.4 Post-Hoc Configuration

The texture descriptor ( $\phi \in \mathbb{R}^{57}$ , Section 3.3) and the fusion procedure are identical for both backbones; only the frozen backbone probabilities differ. The BPR scorer (Eq. (2)) is fit with effective-number class weighting ( $\beta = 0.999$ ) and an asymmetric margin, followed by 1-D logistic calibration. Fusion hyper-parameters are selected per fold on validation AUC over  $\alpha \in \{0.2, 0.3, 0.5\}$ ,  $\gamma \in \{2, 3, 4\}$ , and gate  $\in$

{triangular, entropy}. **Hardware and software.** All experiments were conducted on a workstation running Ubuntu 24.04 LTS with an Intel Core i7-13700K CPU, 32 GB of system memory, and a single NVIDIA GeForce RTX 4090 GPU (24 GB). The software environment was Python with PyTorch (with CUDA/cuDNN) and torchvision for backbone training and inference, timm for backbone components and image transforms, NumPy and SciPy for numerical routines and the paired significance tests, scikit-learn for the evaluation metrics, Pillow for image input/output, and Matplotlib for all plots, including Figs. 3 and 4. The 57-dimensional texture descriptor is implemented directly in NumPy rather than through a radiomics toolkit, so that the GLCM quantization, offset set, and rotation-invariant LBP mapping are fully specified by the source code; the implementation is available at the repository given in the data-availability statement. The post-hoc scorer adds negligible cost relative to backbone training, since it operates on cached backbone probabilities and precomputed descriptors: fitting the scorer, calibrating it, and searching the 18-point fusion grid takes on the order of seconds per fold on a CPU, against hours of GPU time for a single backbone training run.

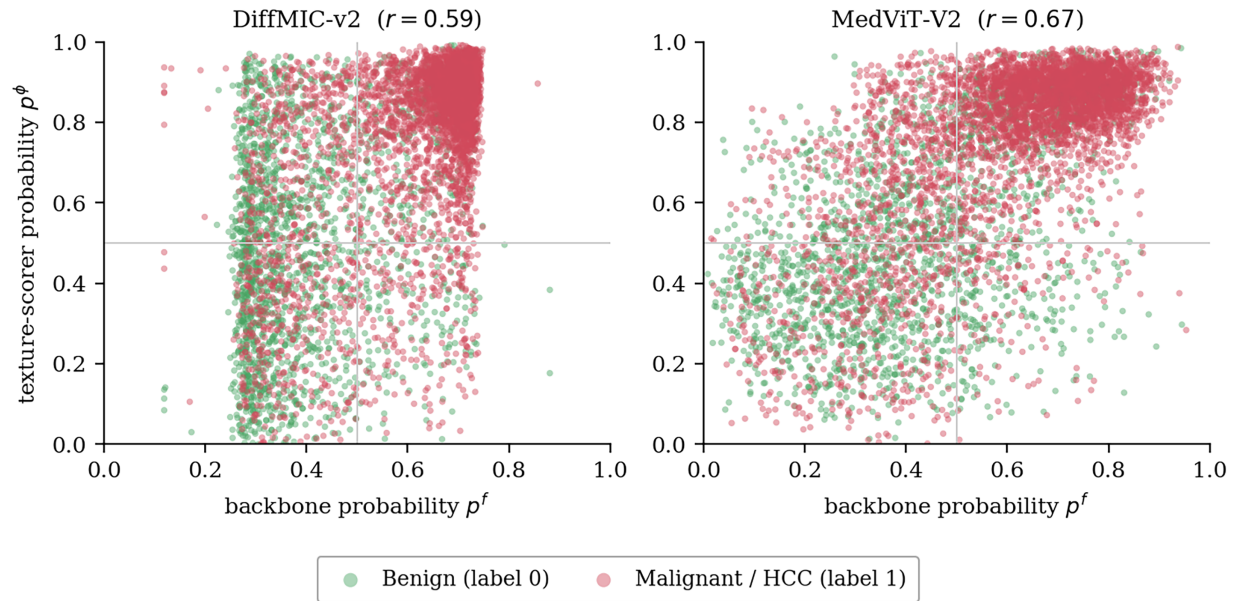


**Figure 4:** Backbone probability  $p^f$  (horizontal axis) vs. texture-scorer probability  $p^\phi$  (vertical axis) for the pooled test predictions of all ten seeds and five folds. Each point is one test image; color denotes the ground-truth label, green for benign (label 0) and red for malignant, i.e., HCC (label 1), as shown in the legend below the panels. The thin gray lines mark the 0.5 decision threshold on each axis. The moderate Pearson correlation in each panel title ( $r = 0.59$  for DiffMIC-v2 and  $r = 0.67$  for MedViT-V2) indicates that the texture axis is complementary rather than redundant.

#### 4.5 Main Results

Table 2 reports both backbones. The diffusion backbone is substantially stronger than the CNN-Transformer baseline (AUC 0.839 vs. 0.739). On top of each frozen backbone, the proposed post-hoc BPR fusion yields statistically significant improvements in sensitivity and F1 on both backbones (DiffMIC-v2: sensitivity +0.0221, F1 +0.0090, both  $p < 0.001$ ; MedViT-V2: sensitivity +0.0455, F1 +0.0269, both  $p < 0.01$ ), together with significant AUC and accuracy gains. The improvement is larger on the weaker MedViT-V2 backbone (AUC +0.0150,  $p < 0.001$ ), where the complementary texture signal contributes more headroom, and smaller but still significant on the already-strong DiffMIC-v2 (AUC +0.0030,  $p < 0.05$ ). The consistent benefit across two heterogeneous backbones suggests that the improvement is not tied to the

diffusion architecture. Fig. 5 makes this pattern visible at a glance: sensitivity, F1, and AUC rise on both backbones—and more strongly on the weaker MedViT-V2—while specificity dips only slightly and remains non-significant on MedViT-V2. The gains come at a small specificity cost, i.e., an operating-point shift toward higher sensitivity; because missing a malignancy is generally the more consequential error in HCC screening, this shift could be advantageous, though its clinical benefit would need to be confirmed in a dedicated study.

**Table 2:** Baseline performance and improvement from post-hoc fusion ( $\Delta$  vs. baseline) for two backbones. Ten-seed repeated 5-fold cross-validation; paired  $t$ -test on the seed-level  $\Delta$ . \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ . BPR is the proposed default scorer; Proto and Hybrid are reported as ablations (Section 4.6). The prototype scorer uses the geometric median.

| Backbone   | Metric      | Baseline (mean $\pm$ SD) | Post-Hoc Fusion ( $\Delta$ vs. Baseline) |            |            |
|------------|-------------|--------------------------|------------------------------------------|------------|------------|
|            |             |                          | BPR (Proposed)                           | Proto      | Hybrid     |
| DiffMIC-v2 | AUC         | 0.8391 $\pm$ 0.0150      | +0.0030*                                 | −0.0002    | +0.0032    |
|            | Accuracy    | 0.7876 $\pm$ 0.0132      | +0.0085**                                | +0.0024    | +0.0052    |
|            | Bal. Acc.   | 0.7719 $\pm$ 0.0115      | +0.0005                                  | −0.0117**  | −0.0035    |
|            | Sensitivity | 0.8143 $\pm$ 0.0370      | +0.0221***                               | +0.0262*** | +0.0200*** |
|            | Specificity | 0.7295 $\pm$ 0.0499      | −0.0210*                                 | −0.0495*** | −0.0270**  |
|            | Precision   | 0.8684 $\pm$ 0.0168      | −0.0060                                  | −0.0168*** | −0.0088*   |
|            | F1          | 0.8397 $\pm$ 0.0141      | +0.0090***                               | +0.0058*   | +0.0066**  |
| MedViT-V2  | AUC         | 0.7389 $\pm$ 0.0158      | +0.0150***                               | +0.0156*** | +0.0169*** |
|            | Accuracy    | 0.7038 $\pm$ 0.0365      | +0.0266**                                | +0.0287**  | +0.0298**  |
|            | Bal. Acc.   | 0.6789 $\pm$ 0.0216      | +0.0155**                                | +0.0171**  | +0.0181**  |
|            | Sensitivity | 0.7462 $\pm$ 0.0791      | +0.0455**                                | +0.0483*** | +0.0497**  |
|            | Specificity | 0.6115 $\pm$ 0.0749      | −0.0145                                  | −0.0140    | −0.0135    |
|            | Precision   | 0.8081 $\pm$ 0.0169      | +0.0032                                  | +0.0040    | +0.0043    |
|            | F1          | 0.7734 $\pm$ 0.0408      | +0.0269**                                | +0.0286**  | +0.0298**  |

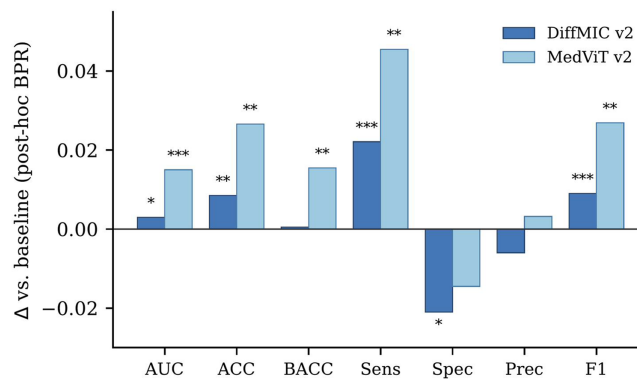


**Figure 5:** Per-metric improvement ( $\Delta$  vs. baseline) of post-hoc BPR fusion for both backbones (BPR columns of Table 2). Stars denote paired  $t$ -test significance (\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ ).

#### 4.6 Ablation: Scorer Choice

The scorer variants in Table 2 reveal a backbone-dependent pattern. On the weaker MedViT-V2, all three scorers behave almost identically and significantly improve the aggregate discrimination metrics

(AUC, accuracy, balanced accuracy, and F1), while their effects on specificity and precision are small and non-significant; the choice of scorer is immaterial, which indicates that the gain originates from the complementary texture representation itself rather than from any particular scoring rule. On the strong DiffMIC-v2, however, the learned BPR scorer is clearly preferable. The non-parametric Proto score over-shifts the operating point toward the malignant class, sharply reducing specificity ( $-0.0495$ ,  $p < 0.001$ ), balanced accuracy ( $-0.0117$ ,  $p < 0.01$ ), and precision ( $-0.0168$ ,  $p < 0.001$ ), and it fails to improve AUC. The BPR scorer regulates this shift: it is the only variant that significantly improves both AUC and accuracy while keeping balanced accuracy neutral, trading a modest specificity decrease for a significant sensitivity and F1 gain. The Hybrid scorer sits between the two. We therefore adopt the class-balanced BPR scorer as the default and report the prototype variants as ablations.

#### 4.7 Ablation: Texture Components

To quantify each texture family’s contribution, we refit the class-balanced BPR scorer on single families (first-order, GLCM, or LBP alone) and on leave-one-out subsets, keeping the rest of the protocol fixed; the “Full (57)” row of Table 3 reproduces the corresponding entries of Table 2. The texture gain is driven mainly by the local binary pattern (LBP) family. On MedViT-V2, LBP alone ( $\Delta\text{AUC} +0.0196$ ) already matches or exceeds the full descriptor ( $+0.0150$ ), and removing LBP roughly halves the AUC gain (to  $+0.0069$ ); on DiffMIC-v2, removing LBP reduces the AUC gain to a negligible  $+0.0006$ . GLCM contributes only a small, secondary signal, and its removal barely changes the aggregate metrics on either backbone. First-order statistics contribute least to discrimination: alone they give no AUC gain on DiffMIC-v2 and only a small one on MedViT-V2, and dropping them from the full set leaves the aggregate metrics essentially unchanged, with slightly higher AUC and F1 on MedViT-V2. The largest specificity costs are incurred by the single-family scorers rather than by first-order statistics specifically (e.g., GLCM alone  $-0.0500$  on MedViT-V2, LBP alone  $-0.0440$  on DiffMIC-v2). Importantly, the sensitivity and F1 gains remain positive and significant for every subset, so the sensitivity-favoring effect is robust to the exact descriptor composition, at a specificity cost that is likewise present throughout. We therefore treat the 57-dimensional descriptor not as an optimum on every metric—a smaller LBP-centric or first-order-free subset performs comparably or slightly better on AUC—but as a compact, conventional, and reproducible default (standard first-order statistics plus two established texture families) that we do not tune per backbone.

**Table 3:** Texture-component ablation. Each row refits the class-balanced BPR scorer on the indicated descriptor subset and re-runs the confidence-gated fusion; entries are  $\Delta$  vs. baseline over ten seeds (paired  $t$ -test; \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ ). “Full (57)” matches the corresponding rows of Table 2.

| Descriptor subset    | DiffMIC-v2         |                     |                     |                   | MedViT-V2          |                     |                     |                   |
|----------------------|--------------------|---------------------|---------------------|-------------------|--------------------|---------------------|---------------------|-------------------|
|                      | $\Delta\text{AUC}$ | $\Delta\text{Sens}$ | $\Delta\text{Spec}$ | $\Delta\text{F1}$ | $\Delta\text{AUC}$ | $\Delta\text{Sens}$ | $\Delta\text{Spec}$ | $\Delta\text{F1}$ |
| Full (57)            | +0.0030*           | +0.0221***          | $-0.0210^*$         | +0.0090***        | +0.0150***         | +0.0455**           | $-0.0145$           | +0.0269**         |
| First-order only (5) | $-0.0002$          | +0.0248***          | $-0.0420^{**}$      | +0.0065*          | +0.0030*           | +0.0324***          | $-0.0310^{***}$     | +0.0153**         |
| GLCM only (16)       | $-0.0026$          | +0.0232***          | $-0.0400^{***}$     | +0.0059**         | +0.0068**          | +0.0499**           | $-0.0500^{***}$     | +0.0228*          |
| LBP only (36)        | +0.0020            | +0.0359***          | $-0.0440^{***}$     | +0.0124***        | +0.0196***         | +0.0460***          | $-0.0125$           | +0.0274**         |
| w/o First-order (52) | +0.0038*           | +0.0232***          | $-0.0295^{**}$      | +0.0079***        | +0.0167***         | +0.0545**           | $-0.0175^*$         | +0.0320**         |
| w/o GLCM (41)        | +0.0028            | +0.0248***          | $-0.0275^{**}$      | +0.0093*          | +0.0166***         | +0.0384***          | $-0.0105$           | +0.0229***        |
| w/o LBP (21)         | +0.0006            | +0.0232***          | $-0.0325^{**}$      | +0.0074*          | +0.0069**          | +0.0471**           | $-0.0390^{***}$     | +0.0234*          |

**Table 3** is best understood as two halves that behave differently, and the contrast between them is informative. Each row is a complete re-run of the pipeline: the class-balanced BPR scorer is refitted from scratch on the indicated descriptor subset, recalibrated, and the fusion hyper-parameters are re-selected on validation AUC, with everything else held fixed; entries are therefore directly comparable to the “Full (57)” row and to each other. Reading the single-family rows against the leave-one-out rows separates two different questions: how much a family can do alone, and how much it adds to the others.

On the weaker MedViT-V2 the two readings agree and point to LBP. LBP alone attains  $\Delta\text{AUC} +0.0196$ , exceeding the full descriptor's  $+0.0150$ , while removing LBP more than halves the gain to  $+0.0069$ ; GLCM alone reaches only  $+0.0068$  and first-order alone  $+0.0030$ . On the stronger DiffMIC-v2 the same conclusion emerges from the leave-one-out column even though the single-family column is uninformative: no single family produces a significant AUC gain in isolation, yet dropping LBP collapses the gain from  $+0.0030$  to a negligible  $+0.0006$ , whereas dropping first-order statistics or GLCM leaves it intact or slightly improves it ( $+0.0038$  and  $+0.0028$ ). This suggests that on an already-strong backbone the texture families are individually too weak to move AUC on their own, and the LBP block is what carries the complementary information once they are combined.

Two further patterns qualify the choice of descriptor. First, the specificity column shows the opposite of what one might expect: the largest specificity losses are incurred by the *narrow* single-family scorers (GLCM alone,  $-0.0500$  on MedViT-V2; LBP alone,  $-0.0440$  on DiffMIC-v2), not by the full descriptor ( $-0.0145$  and  $-0.0210$ ). A scorer built on one family produces a more poorly separated and more poorly calibrated score, which after the confidence gate translates into a larger and less discriminating shift of the operating point; adding the remaining families regularizes the score even where they contribute little discrimination of their own. Second, and for the robustness of the claim, the  $\Delta\text{Sens}$  and  $\Delta\text{F1}$  columns are positive and statistically significant in *every* row of the table, for both backbones and for all seven descriptor subsets. The sensitivity-favoring effect of the method is therefore a property of the fusion mechanism rather than of any particular feature selection, and the descriptor composition determines the size of the effect and its specificity cost rather than its existence.

#### 4.8 ROC/PR Curves, Error Analysis, and Complementarity

**Fig. 6** shows the pooled receiver operating characteristic (ROC) and precision–recall (PR) curves. The post-hoc curve lies on or above the baseline across the operating range for both backbones, with the clearest separation on the weaker MedViT-V2, consistent with the aggregate gains in **Table 2**. **Fig. 7** reports pooled confusion matrices: post-hoc fusion reduces false negatives on both backbones (DiffMIC-v2 808→712, MedViT-V2 1104→906) at a modest increase in false positives, which is exactly the sensitivity-favoring operating-point shift the method is designed to produce. **Fig. 4** plots the backbone probability against the texture-scorer probability; the moderate correlation ( $r = 0.59$  and  $0.67$ ) suggests that the texture descriptor carries partially independent evidence rather than duplicating the backbone, and the off-diagonal cases—confident in one axis but not the other—are where confidence-gated fusion contributes most.

#### 4.9 Selected Fusion Hyper-Parameters

To make the automatic, validation-based hyper-parameter selection transparent, **Table 1** reports how often each fusion configuration was chosen across the 50 folds (ten seeds  $\times$  five folds) per backbone. Both gates are used in practice: the triangular gate is selected in roughly two-thirds of folds (DiffMIC-v2 34/50, MedViT-V2 33/50) and the entropy gate in the remaining third, so neither gate is used exclusively—the choice is made per fold on validation AUC. The smallest mixing weight  $\alpha = 0.2$  is the most frequent choice on both backbones, indicating a conservative texture injection, while the sharpest exponent  $\gamma = 4$  dominates on the

stronger DiffMIC-v2 (31/50), consistent with restricting fusion to the most uncertain, near-boundary cases; on the weaker MedViT-V2 the exponent is distributed more evenly, reflecting that a slightly broader injection is tolerated when the backbone has more headroom.

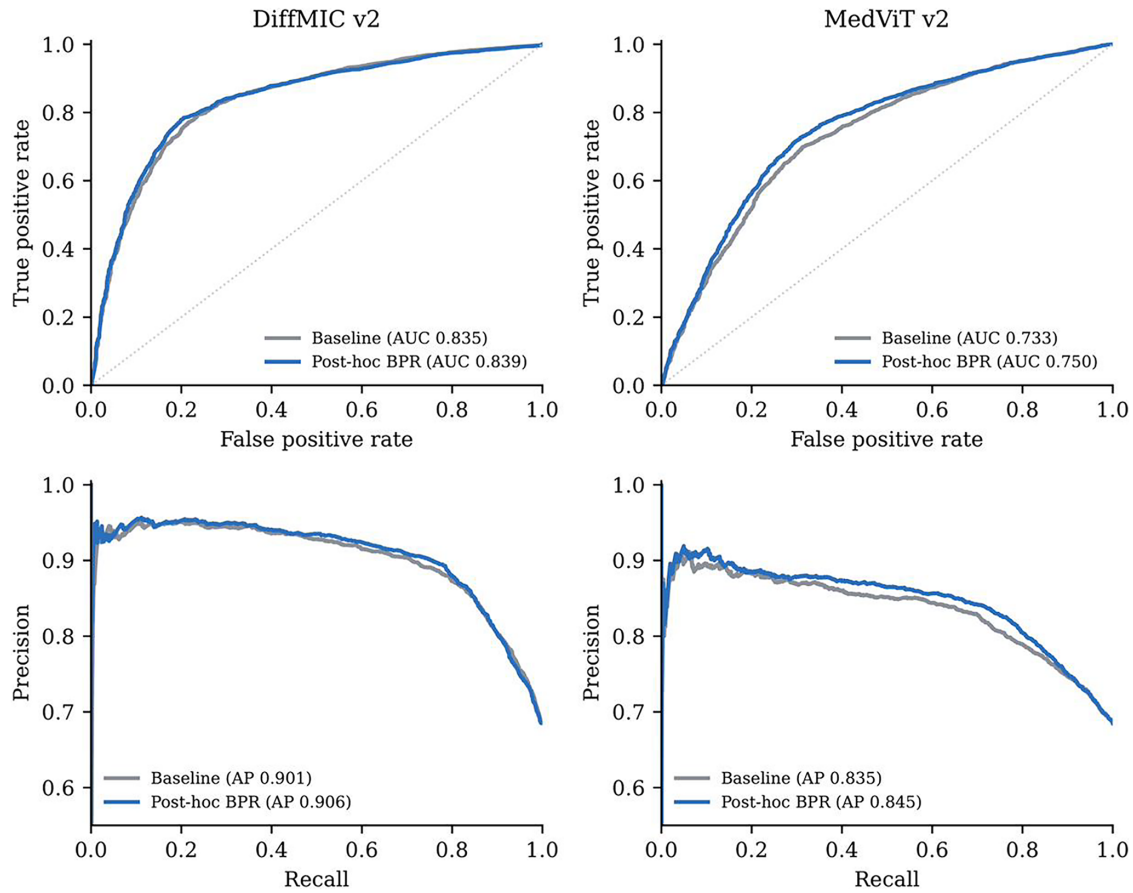


**Figure 6:** Pooled ROC (top) and precision–recall (bottom) curves over all ten seeds and five folds. Post-hoc BPR fusion matches or exceeds the baseline for both backbones. Pooled AUC values differ slightly from the per-seed means in Table 2, because pooling predictions is not equivalent to averaging metrics.

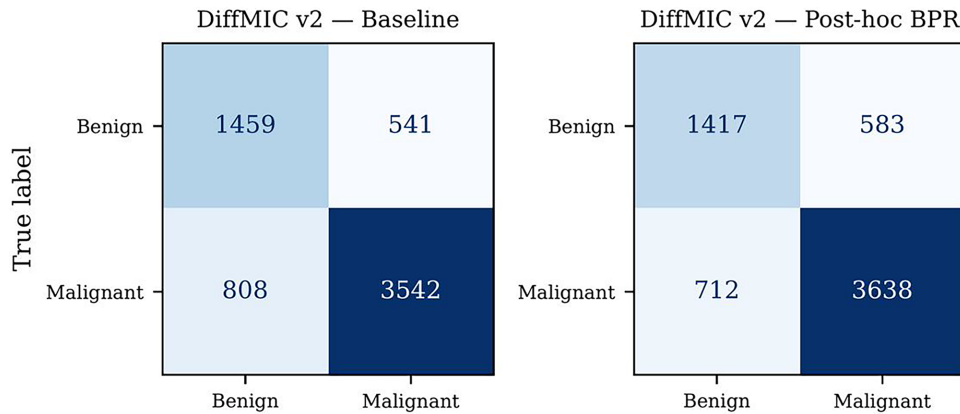


**Figure 7:** (Continued)

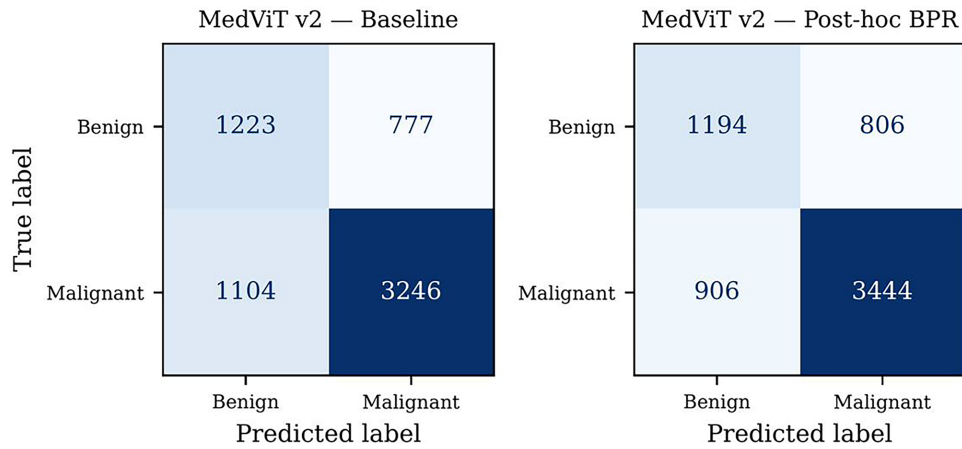


**Figure 7:** Pooled confusion matrices (counts over all folds). Post-hoc fusion lowers false negatives (bold, bottom-left) on both backbones at a small increase in false positives.

#### 4.10 Where the Complementary Signal Helps

To examine *where* the texture signal is useful, and thereby justify confidence-gating directly, we analyze the pooled test predictions inside a near-boundary band  $|p^f - 0.5| < 0.15$  (Table 4). This band covers 34%–41% of cases but concentrates the backbone’s errors. Two observations follow. First, on the weaker MedViT-V2 the texture scorer is more discriminative than the backbone *within* the band (AUC 0.720 vs. 0.641) and corrects more boundary errors than it introduces (54.1% of the 1008 errors flipped correct, against 21.0% of correct cases mislabeled), directly supporting the injection of texture where the backbone is uncertain. Second, on the stronger DiffMIC-v2 the texture scorer is only comparable to the backbone in the band (AUC 0.677 vs. 0.691), and a hard replacement would introduce about as many errors as it fixes (49.3% corrected against 28.4% of a larger correct pool broken). This is precisely why the fusion is *soft* and conservative—a small mixing weight and a ranking scorer that stabilizes the operating point rather than overriding it—and why the DiffMIC-v2 gains are smaller; the net effect of the gated blend remains positive (Fig. 7). Notably, for both backbones the texture scorer has a lower AUC inside the band than in the confident region (0.677/0.720 vs. 0.801/0.780): near-boundary cases are intrinsically hard, so the texture signal is valuable only *relative to the backbone* in that region, and only when blended softly rather than substituted.

**Table 4:** Near-boundary analysis on pooled test predictions (ten seeds  $\times$  five folds;  $N = 6350$  per backbone). The near-boundary band is  $|p^f - 0.5| < 0.15$ ; AUC is computed within each region. “Backbone errors in band” counts cases the backbone misclassifies inside the band, and the indented row reports the share the texture scorer alone labels correctly; the last two rows quantify the opposite risk on backbone-correct cases.

|                                | DiffMIC-v2 | MedViT-V2 |
|--------------------------------|------------|-----------|
| Cases in near-boundary band    | 34.3%      | 40.9%     |
| Backbone AUC (band)            | 0.691      | 0.641     |
| Texture-scorer AUC (band)      | 0.677      | 0.720     |
| Texture-scorer AUC (confident) | 0.801      | 0.780     |
| Backbone errors in band        | 763        | 1008      |

(Continued)

**Table 4 (continued)**

|                          | DiffMIC-v2 | MedViT-V2 |
|--------------------------|------------|-----------|
| Texture correct on them  | 49.3%      | 54.1%     |
| Backbone correct in band | 1412       | 1592      |
| Texture would mislabel   | 28.4%      | 21.0%     |

#### 4.11 Comparison with Existing Approaches

We now position the proposed method against existing work. Two kinds of comparison must be distinguished, because only one of them can be made rigorously on the available evidence.

**Controlled comparison against published state-of-the-art backbones.** The comparison that the experiments support directly is the one already reported in Table 2 and Fig. 5, and it is a controlled one. DiffMIC-v2 [25] and MedViT-V2 [24] are recent, formally published, peer-reviewed medical image classification models with architecturally distinct designs, namely a dual-guidance diffusion classifier and a CNN–Transformer hybrid with Kolmogorov–Arnold networks and dilated neighborhood attention, and they represent the current state of the art for this class of task. Both were trained under our protocol on this dataset and evaluated on exactly the same ten-seed, five-fold partitions as their post-hoc counterparts, so every  $\Delta$  in Table 2 is a strictly paired difference on identical data with identical splits, assessed by a paired  $t$ -test over seed-level differences. Under this comparison the proposed layer significantly improves AUC, accuracy, sensitivity, and F1 over both published baselines. Fig. 5 shows the same result per metric and makes the architecture-dependence visible: the gains are larger on the weaker backbone but present on both, so the improvement is not an artifact of one architecture. It should be noted that these two models serve simultaneously as the state-of-the-art baselines and as the backbones the method operates on; this is inherent to a post-hoc method, whose claim is not that it replaces a strong model but that it improves one without re-training it.

**Why direct numerical comparison with other published results is not appropriate.** One might place our absolute numbers beside those reported elsewhere for ultrasound-based liver lesion classification, but such a comparison would be misleading, and we deliberately avoid it. Published results in this area differ in the dataset, the acquisition equipment, the class definitions, the unit of analysis, and the evaluation protocol, and each of these differences alone can move absolute accuracy by several points. HSQformer [15], the task-specific model released with this dataset, reports strong performance, but the work is an arXiv preprint whose authors state that it is not formally published and that the released code is not organized, with updates pending; adapting it to our exact repeated five-fold protocol would introduce implementation choices that would confound rather than clarify the comparison (Section 4.1). Mitrea et al. [47] report accuracy above 98% by fusing generalized co-occurrence texture features with CNN features on ultrasound, but the task is the discrimination of HCC tissue from surrounding cirrhotic parenchyma on region-level samples from different equipment, which is not the lesion-level benign-vs.-malignant problem addressed here. A large multicenter deep-learning study on B-mode ultrasound [26] operates on a dataset considerably larger than the 635 images available here, in a regime where the data-scarcity constraint that motivates our design does not apply. Reporting these figures side by side would compare datasets rather than methods.

**Method-level comparison.** A more meaningful comparison is with methods that share our operating assumptions, that is, those that modify a trained model's predictions without re-training it. The established family here is post-hoc calibration, of which temperature scaling and Platt scaling are the canonical members [42,56]. These methods rescale the existing decision function monotonically; they can

improve probability quality but cannot reorder samples, and hence cannot change AUC. Our method works differently: it introduces an additional, partially independent source of evidence, which is why it is able to move AUC (Table 2) and to correct a measurable fraction of near-boundary errors (Table 4). Relative to feature-level fusion of hand-crafted and deep descriptors [47], which requires access to and re-training of the classifier that consumes the concatenated features, our method requires only the output probability, and correspondingly claims less: it is a decision-layer adjustment, not a new representation. Within the ablations of Table 2, the class-balanced BPR scorer is also compared against a non-parametric prototype alternative and a hybrid of the two, and it is the only variant that significantly improves both AUC and accuracy on the strong backbone while keeping balanced accuracy neutral.

#### 4.12 Clinical Trade-Off

Across both backbones the proposed method significantly raises sensitivity, AUC, and F1 while incurring only a small specificity cost, and it does so without any backbone re-training. For HCC surveillance—where the dominant cost is generally a missed malignancy—this operating-point shift toward higher sensitivity may be advantageous, subject to clinical validation, and the ability to obtain it post-hoc on an already-deployed frozen model is a practical advantage.

## 5 Discussion

### 5.1 Why a Post-Hoc Intervention

A central empirical observation of this study concerns *where* the errors of a strong backbone lie: on small, noisy ultrasound HCC data the residual errors concentrate in a narrow, near-boundary region rather than being spread across the feature space, and inside that region a simple hand-crafted texture view is competitive with the backbone itself (Section 4.10). It is this localization that a selective, post-hoc intervention is able to exploit. A second, practical consideration is that any training-time intervention requires re-training and re-validating the backbone, which is costly for a model that is already trained and deployed. We stress that the present study does not compare post-hoc fusion with training-time representation-learning objectives and makes no claim about their relative merit; the scope of our evidence is stated explicitly in Section 5.6. Instead of competing with the backbone on the axis it already models well, we introduce a hand-crafted texture axis after training and apply it only where the backbone is uncertain. The post-hoc, backbone-frozen design yields three practical properties. It is model-agnostic, using the identical descriptor and fusion for both a diffusion and a CNN–Transformer backbone. It is leakage-safe, since the scorer, calibration, and hyper-parameters are estimated only on training and validation partitions. Finally, it is inexpensive, fitting a lightweight scorer over cached probabilities and precomputed descriptors instead of repeating a training run.

### 5.2 Why Confidence-Gated Fusion

Uniform averaging of the backbone and texture probabilities degrades confident, correct predictions and collapses specificity, because the texture cue is weaker on easy cases. Confidence gating weights the texture contribution by the backbone's uncertainty, so that confident predictions are preserved and only ambiguous, near-boundary cases are adjusted. This allows the method to raise sensitivity, AUC, and F1 while limiting the specificity cost to a small amount, and it is consistent with the observation that the backbone's errors concentrate near the decision boundary.

### 5.3 Architecture-Dependent Behavior of the Scorer

The scorer ablation shows a backbone-dependent pattern. On the weaker MedViT-V2, all scorer variants improve the aggregate discrimination metrics (AUC, accuracy, balanced accuracy, and F1) almost identically, indicating that the benefit originates from the texture representation itself, not from any particular scoring rule. On the stronger DiffMIC-v2, the learned class-balanced BPR scorer is clearly preferable: the non-parametric prototype score over-shifts the operating point toward the malignant class and sharply reduces specificity, whereas BPR regulates this shift and is the only variant that significantly improves both AUC and accuracy while keeping balanced accuracy neutral. This suggests that when the backbone is already strong, the ordering discipline of a ranking objective is needed to inject the texture signal without distorting the operating point, whereas when the backbone is weak the texture signal helps regardless of how it is scored. In other words, the texture representation and confidence gating are the primary source of the improvement, and the class-balanced ranking scorer acts as a stabilizer that matters most when headroom is small; we therefore position BPR as the principled scorer choice rather than as the sole mechanism behind the gains.

### 5.4 Positioning Relative to Existing Approaches

[Section 4.11](#) sets out the comparison in full; the conclusion we draw from it is worth stating separately here. The evidence supports a claim of the form “this layer improves a given strong backbone under a fixed protocol,” assessed against two recent published models on identical partitions with paired statistics, and it does not support a claim of the form “this system is more accurate than published system X,” because the published systems were evaluated on different data under different protocols. We regard this restraint as appropriate rather than as a weakness of the study: for a post-hoc method the relevant question is the increment it delivers on top of whatever model is already deployed, and that increment is exactly what a paired comparison on shared partitions measures. It also follows that the method is not in competition with the stronger backbones and multicenter systems discussed in [Section 4.11](#); those would be candidate hosts for it. The most informative test of this reading would be to attach the same layer to a model trained on a substantially larger cohort and to observe whether the gain shrinks in proportion to the reduced near-boundary error mass, which is what our analysis predicts.

### 5.5 Clinical Relevance in Ultrasound-Based HCC Detection

Across both backbones the proposed method significantly increases sensitivity, AUC, and F1 at a small specificity cost. In principle, higher sensitivity corresponds to fewer missed malignancies, which could be of particular importance in a disease with poor prognosis such as HCC, where a missed diagnosis may lead to delayed treatment and worse outcomes; establishing this benefit in practice would require clinical evaluation. Ultrasound interpretation is also highly operator-dependent; a model-agnostic decision-support layer that can be attached post-hoc to an already-deployed classifier offers a practical route to more consistent sensitivity without re-validating a new backbone.

### 5.6 Limitations and Future Work

The scope of the experimental evidence should be stated precisely. Every backbone in this study is trained with its original training objective, and every comparison we report, including all ablations, contrasts such a frozen backbone with the same backbone with a post-hoc fusion layer attached. We therefore neither provide nor claim evidence on how post-hoc fusion compares with training-time representation-learning objectives such as ranking, contrastive, or triplet losses applied inside the backbone. Determining whether these objectives would yield smaller, comparable, or larger gains in this data regime (and whether they would be complementary to, rather than substitutes for, post-hoc fusion) requires a dedicated study in which the

objective family, margin, pair-sampling strategy, and optimization schedule are varied systematically, and we regard this as the most direct extension of the present work. Our claims are confined to what the experiments establish: attaching the proposed layer to an already-trained backbone significantly improves sensitivity, AUC, and F1 at a small specificity cost, without touching the backbone.

A second scope limitation concerns the texture representation. The descriptor evaluated here is confined to first-order statistics, second-order GLCM features, and rotation-invariant LBP, and we have not tested the richer multiresolution families reviewed in [Section 2.5](#). Gabor filter-bank responses [45], wavelet-packet sub-band energies, which have already proved effective for liver ultrasound [46], and generalized or superior-order co-occurrence matrices [47] are all plausible substitutes for, or additions to,  $\phi(x)$ . Since the fusion, calibration, and gating machinery treats  $\phi(x)$  as an opaque feature map, such a substitution requires no change to the framework, and the texture ablation of [Section 4.7](#) provides a ready protocol for evaluating it. Whether the additional expressiveness of these families helps or hurts in a 635-image regime is an open question: they carry more parameters and more scale and orientation choices, which is precisely the overfitting risk that motivated our low-capacity default, so we regard this as an empirical question for future work rather than a foregone improvement.

Turning to the data, performance was assessed on a single-center public dataset, which is the only dataset currently accessible for benchmarking HCC ultrasound detection; the absence of multi-center, prospective, and reader-study validation limits external generalizability. The texture descriptor is hand-crafted; while this makes the method deterministic and inexpensive, a learned descriptor might capture additional discriminative structure and is a natural extension. The method also incurs a small, consistent specificity cost as the operating point shifts toward sensitivity; applications that prioritize specificity would require a different gate or selection metric. Finally, method-level design choices (the scorer family and the validation-selection metric) were made on validation data, following standard practice, but a fully prespecified, multi-center evaluation would provide stronger evidence. Future work will pursue learned descriptors, multi-center validation as data become available, and integration with explainability tooling to support human oversight [58].

## 6 Conclusion

We proposed a post-hoc, backbone-frozen texture–ranking fusion framework for data-efficient HCC detection from ultrasound images. Motivated by the observation that the residual errors of strong backbones concentrate in a narrow region near the decision boundary, and by the practical value of an intervention that requires no backbone re-training, the method injects a hand-crafted texture descriptor through a class-balanced BPR ranking scorer and a confidence-gated fusion rule, using only training and validation data so that the evaluation remains leakage-safe.

Under ten-seed repeated 5-fold cross-validation on two architecturally distinct backbones, the method delivered statistically significant and consistent gains. On the weaker CNN–Transformer MedViT-V2 it raised AUC from 0.7389 to 0.7539 ( $\Delta = +0.0150$ ,  $p < 0.001$ ), sensitivity from 0.7462 to 0.7917 ( $\Delta = +0.0455$ ,  $p < 0.01$ ), F1 from 0.7734 to 0.8003 ( $\Delta = +0.0269$ ,  $p < 0.01$ ), and balanced accuracy by  $+0.0155$  ( $p < 0.01$ ), at a specificity cost of 0.0145 that was not statistically significant. On the already-strong diffusion classifier DiffMIC-v2 it raised AUC from 0.8391 to 0.8421 ( $\Delta = +0.0030$ ,  $p < 0.05$ ), sensitivity from 0.8143 to 0.8364 ( $\Delta = +0.0221$ ,  $p < 0.001$ ), and F1 from 0.8397 to 0.8487 ( $\Delta = +0.0090$ ,  $p < 0.001$ ), for a specificity cost of 0.0210 and no loss of balanced accuracy ( $\Delta = +0.0005$ , not significant). In the pooled confusion matrices over all 50 test partitions, false negatives fell from 808 to 712 on DiffMIC-v2 and from 1104 to 906 on MedViT-V2, that is, by 11.9% and 17.9%, respectively. The mechanism analysis localizes these gains: the near-boundary band  $|p^f - 0.5| < 0.15$  contains 34.3% and 40.9% of test cases but concentrates the errors, and

within it the texture scorer alone labels 49.3% and 54.1% of the backbone's mistakes correctly. All of this was obtained with a mixing weight bounded by  $\alpha \leq 0.5$  and selected as 0.2 in roughly half of all folds, and without a single gradient step through either backbone. Ablations indicate that these gains stem chiefly from the texture axis—and, within it, mainly from local binary pattern features—and confidence gating, with the class-balanced ranking scorer providing stability—most visibly on the stronger backbone, where it curbs the specificity loss incurred by a non-parametric alternative. This consistency across heterogeneous backbones suggests that the approach is model-agnostic, at least across the two backbones evaluated here. Future work will focus on learned descriptors and on multi-center HCC ultrasound validation as such datasets become publicly accessible.

**Acknowledgement:** Hosoo Shin and Seongyeon Son contributed equally to this work.

**Funding Statement:** This research was supported by the Keimyung Research Initiative of Keimyung University.

**Author Contributions:** The authors confirm contribution to the paper as follows: conceptualization, Hosoo Shin and Eunchan Kim; methodology, Hosoo Shin, Seongyeon Son and Minhee Park; software, Hosoo Shin, Seongyeon Son and Minhee Park; validation, Hosoo Shin, Seongyeon Son, Minhee Park and Yonghyun Lee; formal analysis, Hosoo Shin; investigation, Hosoo Shin, Seongyeon Son, Minhee Park, Lin Xia and Yonghyun Lee; data curation, Hosoo Shin, Seongyeon Son and Minhee Park; writing—original draft preparation, Hosoo Shin; writing—review and editing, Seongyeon Son, Minhee Park, Lin Xia, Yonghyun Lee and Eunchan Kim; visualization, Hosoo Shin, Seongyeon Son, Minhee Park and Lin Xia; supervision, Eunchan Kim; project administration, Eunchan Kim; funding acquisition, Yonghyun Lee. All authors reviewed and approved the final version of the manuscript.

**Availability of Data and Materials:** This study used a publicly available liver ultrasound dataset provided by the Third Affiliated Hospital of Sun Yat-sen University, released with the associated preprint [15] at <https://github.com/Asunatan/HSQformer>. The implementation of the proposed method and the experimental code are openly available at [https://github.com/shinhosoo/bpr\\_hcc](https://github.com/shinhosoo/bpr_hcc).

**Ethics Approval:** This study is a secondary analysis of de-identified data and involved no new human or animal experimentation. All experiments used the liver ultrasound images released with the retrospective study of She et al. [15] (<https://github.com/Asunatan/HSQformer/tree/main/dataset>), which constitute the single-center test cohort of that study and consist of B-mode images acquired at the Third Affiliated Hospital of Sun Yat-sen University. The originating study was retrospective and reports that informed consent was not required because fully anonymized images were used [15]. According to the documentation accompanying the public release, the released data received ethical certification from the Third Affiliated Hospital of Sun Yat-sen University on 15 May 2023. As described by She et al. [15], the original DICOM images were losslessly converted to PNG, and textual metadata, including patient identifiers, device information, institutional details, and date stamps, were removed by automated detection. The dataset is distributed for academic research use only and was used in accordance with those terms. The authors had no access to identifiable patient information at any stage.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Appendix A Additional Details

### Appendix A.1 Texture Descriptor

The complementary descriptor  $\phi(x) \in \mathbb{R}^{57}$  (Section 3.3) is fully deterministic and requires no training. It concatenates: (i) five first-order intensity statistics (mean, standard deviation, skewness, kurtosis, entropy); (ii) sixteen GLCM features—contrast, homogeneity, energy, and correlation, each at four offsets  $\{(0, 1), (1, 0), (1, 1), (1, -1)\}$  with  $L = 16$  quantized gray levels; and (iii) a 36-bin rotation-invariant local

binary pattern histogram. The descriptor is standardized per fold using statistics computed on the training partition only, and the same implementation is applied to both backbones.

### ***Appendix A.2 Hyper-Parameter Grid and Selection***

For every fold the fusion hyper-parameters are selected by maximizing validation AUC over the grid  $\alpha \in \{0.2, 0.3, 0.5\}$ ,  $\gamma \in \{2, 3, 4\}$ , and  $\text{gate} \in \{\text{triangular}, \text{entropy}\}$  (18 combinations), with the scorer and its logistic calibration held fixed. The BPR scorer (Eq. (2)) uses an asymmetric margin and effective-number class weighting ( $\beta = 0.999$ ). No test labels are consulted at any point in scorer fitting, calibration, or hyper-parameter selection.

### ***Appendix A.3 Paired Evaluation and Leakage Safety***

At each of the ten seeds the dataset is independently re-partitioned into a 5-fold split; a backbone and its post-hoc counterpart share identical partitions at every seed and fold, so the reported  $\Delta$  values are strictly paired. Significance is assessed with a paired  $t$ -test over the ten seed-level differences. Because scorer fitting, calibration, and hyper-parameter selection use only the training and validation partitions, the test partition functions as a genuinely held-out set for final evaluation.

## **References**

1. Llovet JM, Kelley RK, Villanueva A, Singal AG, Pikarsky E, Roayaie S, et al. Hepatocellular carcinoma. *Nat Rev Dis Primers*. 2021;7(1):6. doi:10.1038/s41572-020-00240-3.
2. Vogel A, Meyer T, Sapisochin G, Salem R, Saborowski A. Hepatocellular carcinoma. *Lancet*. 2022;400(10360):1345–62. doi:10.1016/S0140-6736(22)01200-4.
3. Singal AG, Zhang E, Narasimman M, Rich NE, Waljee AK, Hoshida Y, et al. HCC surveillance improves early detection, curative treatment receipt, and survival in patients with cirrhosis: a meta-analysis. *J Hepatol*. 2022;77(1):128–39. doi:10.1016/j.jhep.2022.01.023.
4. Hamza MN, Alibakhshikenari M, Islam MT, Lavadiya S, ud Din I, Sanches B, et al. High-sensitivity nanomaterial near-infrared biosensor for label-free early cancer detection via exosomal biomarkers. *Appl Opt*. 2026;65(1):39–54. doi:10.1364/AO.581804.
5. Hamza MN, Islam MT, Lavadiya S, ud Din I, Sanches B, Koziel S, et al. A 50THz ultra-wideband nano-photonics perfect absorber biosensor for label-free detection of circulating cancer exosomes: advancing early cancer diagnostics. *IEEE Photonics J*. 2025;17(4):5900316. doi:10.1109/JPHOT.2025.3571811.
6. Tanaka H. Current role of ultrasound in the diagnosis of hepatocellular carcinoma. *J Med Ultrason*. 2020;47(2):239–55. doi:10.1007/s10396-020-01012-y.
7. Singal A, Volk ML, Waljee A, Salgia R, Higgins P, Rogers MAM, et al. Meta-analysis: surveillance with ultrasound for early-stage hepatocellular carcinoma in patients with cirrhosis. *Aliment Pharmacol Ther*. 2009;30(1):37–47. doi:10.1111/j.1365-2036.2009.04014.x.
8. Tzartzeva K, Obi J, Rich NE, Parikh ND, Marrero JA, Yopp A, et al. Surveillance imaging and alpha fetoprotein for early detection of hepatocellular carcinoma in patients with cirrhosis: a meta-analysis. *Gastroenterology*. 2018;154(6):1706–18. doi:10.1053/j.gastro.2018.01.064.
9. Kim SY, An J, Lim YS, Han S, Lee JY, Byun JH, et al. MRI with liver-specific contrast for surveillance of patients with cirrhosis at high risk of hepatocellular carcinoma. *JAMA Oncol*. 2017;3(4):456–63. doi:10.1001/jamaoncol.2016.3147.
10. Hu H, Zhao Y, He C, Qian L, Huang P. Ultrasonography of hepatocellular carcinoma: from diagnosis to prognosis. *J Clin Transl Hepatol*. 2024;12(5):516–24. doi:10.14218/JCTH.2024.00018.
11. Wu K, Chen X, Ding M. Deep learning based classification of focal liver lesions with contrast-enhanced ultrasound. *Optik*. 2014;125(15):4057–63. doi:10.1016/j.ijleo.2014.01.114.

12. Chaiteerakij R, Ariyaskul D, Kulkraisri K, Apiparakoon T, Sukcharoen S, Chaichuen O, et al. Artificial intelligence for ultrasonographic detection and diagnosis of hepatocellular carcinoma and cholangiocarcinoma. *Sci Rep.* 2024;14(1):20617. doi:10.1038/s41598-024-71657-z.
13. Dana J, Meyer A, Paisant A, Rode A, Sartoris R, S  r  r O, et al. Improving risk stratification and detection of early HCC using ultrasound-based deep learning models. *JHEP Rep.* 2025;7(10):101510. doi:10.1016/j.jhepr.2025.101510.
14. Song K, Feng J, Chen D. A survey on deep learning in medical ultrasound imaging. *Front Phys.* 2024;12:1398393. doi:10.3389/fphy.2024.1398393.
15. She C, Lu R, He D, Lv J, Lin Y, Cheng M, et al. A retrospective systematic study on hierarchical sparse query transformer-assisted ultrasound screening for early hepatocellular carcinoma. *arXiv:2502.03772.* 2025.
16. Lu RF, She CY, He DN, Cheng MQ, Wang Y, Huang H, et al. AI enhanced diagnostic accuracy and workload reduction in hepatocellular carcinoma screening. *npj Digit Med.* 2025;8(1):500. doi:10.1038/s41746-025-01892-9.
17. Rhyou SY, Yoo JC. Automated ultrasonography of hepatocellular carcinoma using discrete wavelet transform based deep-learning neural network. *Med Image Anal.* 2025;101(10):103453. doi:10.1016/j.media.2025.103453.
18. Zhao Z, Alzubaidi L, Zhang J, Duan Y, Naseem U, Gu Y. Robust and explainable framework to address data scarcity in diagnostic imaging. *Comput Biol Med.* 2025;197(6):111052. doi:10.1016/j.compbiomed.2025.111052.
19. Candemir S, Nguyen XV, Folio LR, Prevedello LM. Training strategies for radiology deep learning models in data-limited scenarios. *Radiol Artif Intell.* 2021;3(6):e210014. doi:10.1148/ryai.2021210014.
20. Mbonihankuye S, Nkunzimana A, Ndagijimana A. Healthcare data security technology: HIPAA compliance. *Wirel Commun Mob Comput.* 2019;2019(1):1927495. doi:10.1155/2019/1927495.
21. Singh AK, Anand A, Lv Z, Ko H, Mohan A. A survey on healthcare data: a security perspective. *ACM Trans Multimed Comput Commun Appl.* 2021;17(2s):59. doi:10.1145/3422816.
22. Dhar T, Dey N, Borra S, Sherratt RS. Challenges of deep learning in medical image analysis—improving explainability and trust. *IEEE Trans Technol Soc.* 2023;4(1):68–75. doi:10.1109/tts.2023.3234203.
23. Vega R, Dehghan M, Nagdev A, Buchanan B, Kapur J, Jaremko JL, et al. Overcoming barriers in the use of artificial intelligence in point of care ultrasound. *npj Digit Med.* 2025;8(1):213. doi:10.1038/s41746-025-01633-y.
24. Manzari ON, Asgariandehkordi H, Koleilat T, Xiao Y, Rivaz H. Medical image classification with KAN-integrated transformers and dilated neighborhood attention. *Appl Soft Comput.* 2025;186(11):114045. doi:10.1016/j.asoc.2025.114045.
25. Yang Y, Fu H, Aviles-Rivero AI, Xing Z, Zhu L. DiffMIC-v2: medical image classification via improved diffusion network. *IEEE Trans Med Imag.* 2025;44(5):2244–55. doi:10.1109/TMI.2025.3530399.
26. Dadoun H, Rousseau AL, de Kerviler E, Corre  as JM, Tissier AM, Joujou F, et al. Deep learning for the detection, localization, and characterization of focal liver lesions on abdominal US images. *Radiol Artif Intell.* 2022;4(3):e210110. doi:10.1148/ryai.210110.
27. Calderaro J, Seraphin TP, Luedde T, Simon TG. Artificial intelligence for the prevention and clinical management of hepatocellular carcinoma. *J Hepatol.* 2022;76(6):1348–61. doi:10.1016/j.jhep.2022.01.014.
28. Masokano IB, Liu W, Xie S, Marcellin DFH, Pei Y, Li W. The application of texture quantification in hepatocellular carcinoma using CT and MRI: a review of perspectives and challenges. *Cancer Imag.* 2020;20(1):67. doi:10.1186/s40644-020-00341-y.
29. Su Z, Guo J, Yang X, Wang Q, Coenen F, Huang K. Navigating distribution shifts in medical image analysis: a survey. *arXiv:2411.05824.* 2024.
30. Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping. In: *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*; 2006 Jun 17–22; New York, NY, USA. p. 1735–42. doi:10.1109/CVPR.2006.100.
31. Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering. In: *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*; 2015 Jun 7–12; Boston, MA, USA. p. 815–23. doi:10.1109/CVPR.2015.7298682.
32. Wen Y, Zhang K, Li Z, Qiao Y. A discriminative feature learning approach for deep face recognition. In: *Computer Vision–ECCV 2016 (ECCV 2016)*. Cham, Switzerland: Springer; 2016. p. 499–515. doi:10.1007/978-3-319-46478-7\_31.

33. Movshovitz-Attias Y, Toshev A, Leung TK, Ioffe S, Singh S. No fuss distance metric learning using proxies. In: Proceedings of the 2017 IEEE International Conference on Computer Vision; 2017 Oct 22–29; Venice, Italy. p. 360–8. doi:10.1109/ICCV.2017.47.
34. Nguyen H, Nguyen H, Chang M, Pham H, Narayanan S, Pazzani M. ConPro: learning severity representation for medical images using contrastive learning and preference optimization. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2024 Jun 17–18; Seattle, WA, USA. p. 5105–12.
35. Mazaheri P, Bidgoli AA, Rahnamayan S, Tizhoosh HR. Ranking loss and sequestering learning for reducing image search bias in histopathology. *Appl Soft Comput*. 2023;142(12):110346. doi:10.1016/j.asoc.2023.110346.
36. Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L. BPR: Bayesian personalized ranking from implicit feedback. In: Proceedings of the 2009 Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI); 2009 Jun 18–21; Montreal, QC, Canada. p. 452–61.
37. Peska L, Buza K, Koller J. Drug-target interaction prediction: a Bayesian ranking approach. *Comput Methods Programs Biomed*. 2017;152(80):15–21. doi:10.1016/j.cmpb.2017.09.003.
38. Lan W, Chen W, Li C, Chen Q, Peng W, Chen YPP, et al. CLTDA: identifying associations between tsRNAs and diseases based on contrastive learning. *Big Data Min Anal*. 2025;8(6):1324–34. doi:10.26599/BDMA.2025.9020027.
39. Hu X, Zhang C, Fan Y, Wu K, Deng L. AGDNGDA: unraveling drug-associated genes with adaptive graph diffusion networks. *J Chem Inf Model*. 2025;65(22):12619–26. doi:10.1021/acs.jcim.5c02276.
40. Song J, Gong Y, Song Z, Xu X, Qian K, Liu Y. DriverSub-SVM: a machine learning approach for cancer subtype classification by integrating patient-specific and global driver genes. *BMC Bioinform*. 2025;26(1):297. doi:10.1186/s12859-025-06318-7.
41. Zheng K, Wang J, Xu J. Graph-theoretic insights into Bayesian personalized ranking for recommendation. In: Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025); 2025 Dec 2–7; San Diego, CA, USA.
42. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML); 2017 Aug 6–11; Sydney, Australia. p. 1321–30.
43. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern*. 1973;SMC-3(6):610–21. doi:10.1109/TSMC.1973.4309314.
44. Ojala T, Pietikäinen M, Mäenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell*. 2002;24(7):971–87. doi:10.1109/TPAMI.2002.1017623.
45. Manjunath BS, Ma WY. Texture features for browsing and retrieval of image data. *IEEE Trans Pattern Anal Mach Intell*. 1996;18(8):837–42. doi:10.1109/34.531803.
46. Virmani J, Kumar V, Kalra N, Khandelwal N. SVM-based characterization of liver ultrasound images using wavelet packet texture descriptors. *J Digit Imag*. 2013;26(3):530–43. doi:10.1007/s10278-012-9537-8.
47. Mitrea DA, Brehar R, Nedeveschi S, Lupșor-Platon M, Socaciu M, Badea R. Hepatocellular carcinoma recognition from ultrasound images using combinations of conventional and deep learning techniques. *Sensors*. 2023;23(5):2520. doi:10.3390/s23052520.
48. Kumar D, Bakariya B, Verma C, Illés Z. LivXAI-Net: an explainable AI framework for liver disease diagnosis with IoT-based real-time monitoring support. *Comput Methods Programs Biomed*. 2025;270(4):108950. doi:10.1016/j.cmpb.2025.108950.
49. Dimopoulos MP, Mulita F, Spyridonidis I, Spyridonidis T, Kitrou P, Karnabatidis D, et al. Hepatocolic fistula after ablative radiation lobectomy of a right lobe colorectal liver metastasis. *Gastroenterol Rev*. 2025;20(4):461–4. doi:10.5114/pg.2025.156469.
50. Dritsas S, Frountzas M, Schizas D, Mulita F, Katsimpoulas M, Gkolfakis P, et al. The role of Toll-like receptors in liver regeneration after partial hepatectomy. *Gastroenterol Rev*. 2025;20(3):294–304. doi:10.5114/pg.2025.154724.
51. Sun Y, Jin G. Primary squamous cell carcinoma of the liver: a case report. *J Int Med Res*. 2021;49(6):3000605211021275. doi:10.1177/03000605211021275.
52. Song Y, Shi J, Zhang X, Qiao M, Sun Z, Tian S. Diagnostic value of imaging modalities in primary squamous cell carcinoma of the liver. *J Clin Ultrasound*. 2023;51(5):887–97. doi:10.1002/jcu.23447.

53. Igami T, Nakamura Y, Hirose T, Ebata T, Yokoyama Y, Sugawara G, et al. Application of a three-dimensional print of a liver in hepatectomy for small tumors invisible by intraoperative ultrasonography: preliminary experience. *World J Surg.* 2014;38(12):3163–6. doi:10.1007/s00268-014-2740-7.
54. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology.* 2020;295(2):328–38. doi:10.1148/radiol.2020191145.
55. Cui Y, Jia M, Lin TY, Song Y, Belongie S. Class-balanced loss based on effective number of samples. In: *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2019 Jun 15–20; Long Beach, CA, USA. p. 9260–9. doi:10.1109/CVPR.2019.00949.
56. Platt JC. Probabilities for SV machines. In: Smola AJ, Bartlett P, Schölkopf B, Schuurmans D, editors. *Advances in large-margin classifiers*. Cambridge, MA, USA: MIT Press; 2000. p. 61–74. doi:10.7551/mitpress/1113.003.0008.
57. Vardi Y, Zhang CH. The multivariate LI-median and associated data depth. *Proc Natl Acad Sci U S A.* 2000;97(4):1423–6. doi:10.1073/pnas.97.4.1423.
58. Kotter E, D'Antonoli TA, Cuocolo R, Hierath M, Huisman M, Klontzas ME, et al. Guiding AI in radiology: ESR's recommendations for effective implementation of the European AI Act. *Insights Imag.* 2025;16(1):33. doi:10.1186/s13244-025-01905-x.