Learning Latent Imaging Biomarkers for Interpretable Microvascular Invasion Prediction in Hepatocellular Carcinoma

Ji Rao[1], Xinyu Liu[1], Yong Yi[2], Ying Xiao[1], Ye Luo[1]*

[1]School of Computer Science and Technology, Tongji University
[2]Department of Liver Surgery & Transplantation, Zhongshan Hospital, Fudan University
yeluo@tongji.edu.cn

Abstract

Microvascular invasion (MVI) is a critical prognostic factor that significantly impacts postoperative outcomes in hepatocellular carcinoma (HCC). As the current gold standard for the diagnosis of MVI is based on the postoperative histopathological examination of whole slide images, accurate preoperative prediction of MVI status using magnetic resonance imaging (MRI) presents both a substantial clinical imperative and a significant challenge. In order to discover reliable MRI-based imaging biomarkers to support clinical decision making and enhance the interpretability of deep learning-based diagnostic models, we propose a novel interpretable MVI prediction framework in which the shared latent visual attributes are first learned and then used for potential imaging biomarker extraction and MVI diagnosis, respectively. To ensure that the visual attributes of these biomarkers are generalizable across diverse patients, the similarity constraints at the intra-patient level and the inter-patient level are enforced within the learned feature space, enabling intuitive biomarker discovery directly from the original image space. To guarantee semantic alignment between biomarkers and the characteristics of individual patients, we introduce a novel classification mechanism that directly links the alignment between each biomarker and patient-specific characteristics with the prediction, thereby ensuring a precise prediction of MVI. Furthermore, the interpretability of the model is enhanced by integrating a mask-based visual explanation method that highlights regions in patient images that correspond to the identified biomarkers. Extensive experiments on two MVI prediction datasets: HCC-WCH and HCC-ZSH unequivocally demonstrate our method's superior performance in both classification accuracy and interpretability.

Code — https://github.com/Ross-Rao/MVI-IBs

1 Introduction

Hepatocellular Carcinoma (HCC) ranks as the sixth most common cancer globally and the third leading cause of cancer-related deaths, with a particularly severe impact in the Asia-Pacific region (Mak et al. 2024), thus accurate preoperative prediction of HCC holds significant importance for guiding patient treatment strategies and prognostic evaluation. The presence of Microvascular Invasion (MVI) is strongly associated with aggressive tumor biological behavior and poor prognosis, representing a primary risk factor influencing post-resection survival outcomes and intrahepatic metastasis in HCC patients (Wang et al. 2021). Currently, MVI diagnosis predominantly relies on pathological analysis of patient tissue images by pathologists after surgeries (Yao et al. 2023; Zhang et al. 2023). Consequently, developing less invasive preoperative MVI diagnostic methods is crucial in the field of HCC.

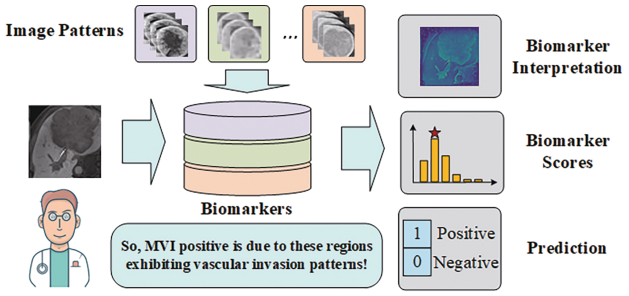

Figure 1: By learning image patterns, our proposed IRCL model provides imaging biomarkers as explanatory evidence during MVI classification, thereby yielding novel medical knowledge and offering clinicians multi-faceted evidence for understanding model decisions.

Recently, various deep learning methods have been proposed to analyze preoperative MRI images for MVI prediction (Liu, Yi, and Luo 2024; Zhang et al. 2024), aiming to alleviate radiologists' workload while simultaneously improving diagnosis precision and efficiency. However, due to the intricate coupling and non-linear entanglement between deep learning based features, experts and clinicians often find it challenging to understand which specific microscopic features or phenotypes determine or dominate the model's predictions. Although some interpretability methods have been proposed, such as identifying result-relevant regions through patch selection (Zhao et al. 2022) or generating heatmaps for direct explanation (Pang et al. 2025; Zheng, Yi, and Luo 2025), these approaches typically lack control over the specific features learned by the model. Furthermore, they cannot guarantee whether the identified features are generalizable across different patients or if they

*Corresponding author

genuinely contribute to the prediction decisions.

Meanwhile, imaging biomarkers, as a widely utilized tool in traditional radiomics for disease diagnosis, efficacy assessment, and treatment planning, have been extensively applied in both scientific research and clinical practice (O'connor et al. 2017). However, current efforts in discovering imaging biomarkers from MRI for MVI remain limited. Furthermore, the high-dimensional features extracted by deep learning models often lack intuitive interpretability, posing a challenge in understanding the specific attribution regions of these biomarkers within patient images (Rotem et al. 2024). Therefore, as illustrated in Fig. 1, there is a pressing need to develop methods capable of explaining the precise regions of interest of imaging biomarkers in patient images, thereby generating novel knowledge for clinical prognosis.

In order to address the aforementioned issues, we propose a novel two-stage interpretable MVI prediction framework, termed **Interpretable Representation Contrastive Learning (IRCL)**. Stage I of this framework leverages unsupervised learning to capture the intra-patient and inter-patient features, which are then clustered into imaging biomarkers; Stage II subsequently utilizes these biomarkers for MVI prediction. Specifically, in the first stage, a latent representation learning network is built to map the raw image data into a latent representation space; thus, the completeness of the original image information and the interpretability of the latent representation in terms of the visual attributes can be endowed simultaneously. To further optimize these latent representations, a contrastive learning approach is proposed by imposing the similarity constraints at both the intra-patient level and the inter-patient level. This dual-level constraint enables the model to learn the intrinsic distribution of the entire dataset. Subsequently, the obtained latent representations are clustered to identify and extract latent imaging biomarkers that represent common patterns and crucial features for MVI diagnosis. In the second stage, we perform Transformer-based prediction by leveraging the latent imaging biomarkers extracted in the first stage, in conjunction with individual patient characteristics. Specifically, the core of this stage involves evaluating the degree of alignment between individual patient features and the identified candidate latent imaging biomarkers. Furthermore, we design a mask-based visual interpretability method. This approach optimizes an objective function to compel features of masked images to approximate the corresponding imaging biomarkers, thereby constraining the mask to identify the projection regions of these biomarkers in the original image, serving as their interpretability method.

Our main contributions can be summarized as follows:

- We propose an MVI biomarker mining network via aggregating intra-patient and inter-patient features to obtain high-dimensional MVI imaging biomarkers, enabling MVI prediction based on the association between these biomarkers and patient characteristics.
- A mask-based biomarker interpretation method is designed to effectively highlight the projection regions of imaging biomarkers within the original images through optimized constraints.
- Extensive experiments on two MVI datasets demonstrate the leading predictive and interpret-ability performance.

2 Related Work

Multi-Modal MVI Classification. Methods for MVI diagnosis based on MRI images can be primarily categorized into three groups: radiomics approaches (Zhou et al. 2024), Convolutional Neural Network (CNN)-based methods, and Transformer-based methods. (Zheng et al. 2025) developed a topology-aware deep learning model leveraging MRI, which emphasized topological features for MVI prediction and prognostic stratification. (Zhang et al. 2024) employed knowledge distillation to integrate clinical data and multi-modal MRI images, thereby enhancing HCC classification performance using only MRI images. (Pang et al. 2025) proposed a dual-branch deep multiple instance learning framework that extracts 2D and 3D image features via Hierarchical Attention and Consistency Learning modules, respectively. The consistency learning module further aligns the attention weights of the two branches, leading to improved classification performance. (Wang et al. 2024) introduced a multi-task deep learning model based on the Transformer architecture, capable of simultaneously predicting MVI and recurrence-free survival. (Liu, Yi, and Luo 2024) utilized fine-grained features from multi-modal MRI images, employing a Multi-modal Fine-Grained generator and an Attention Pooling module to extract and fuse features from different modalities, thereby enhancing the accuracy of MVI grading.

DL-based Biomarker Identification. Deep learning has emerged as a transformative approach for biomarker identification across diverse biomedical domains, leveraging its ability to extract complex patterns from high-dimensional data. (Liang et al. 2023) developed PathFinder, a human-centric DL framework for discovering tissue biomarkers in liver cancer prognosis, identifying the spatial distribution of necrosis as a critical prognostic biomarker, proposing necrosis area fraction and tumor necrosis distribution as clinically independent indicators. (Zhao et al. 2022) proposes a novel prognostic imaging biomarker discovery method for survival analysis in idiopathic pulmonary fibrosis. The approach learns local region representations from lung images via a contrastive learning model, then groups similar imaging patterns using a clustering method. Subsequently, an efficient Clustering Vision Transformer aggregates these local region representations to predict patient mortality risk and identify high-risk patterns.

3 The Proposed Method

3.1 Overall Architecture

Fig. 2 describes the overall architecture of Interpretable Representation Contrastive Learning (IRCL). Our proposed method comprises two pivotal stages. In stage I, we employ a cascade encoder-decoder architecture to extract latent representations from images. These representations are then reconstructed into images, serving as a spatial regularization constraint. To ensure generalizability of the visual

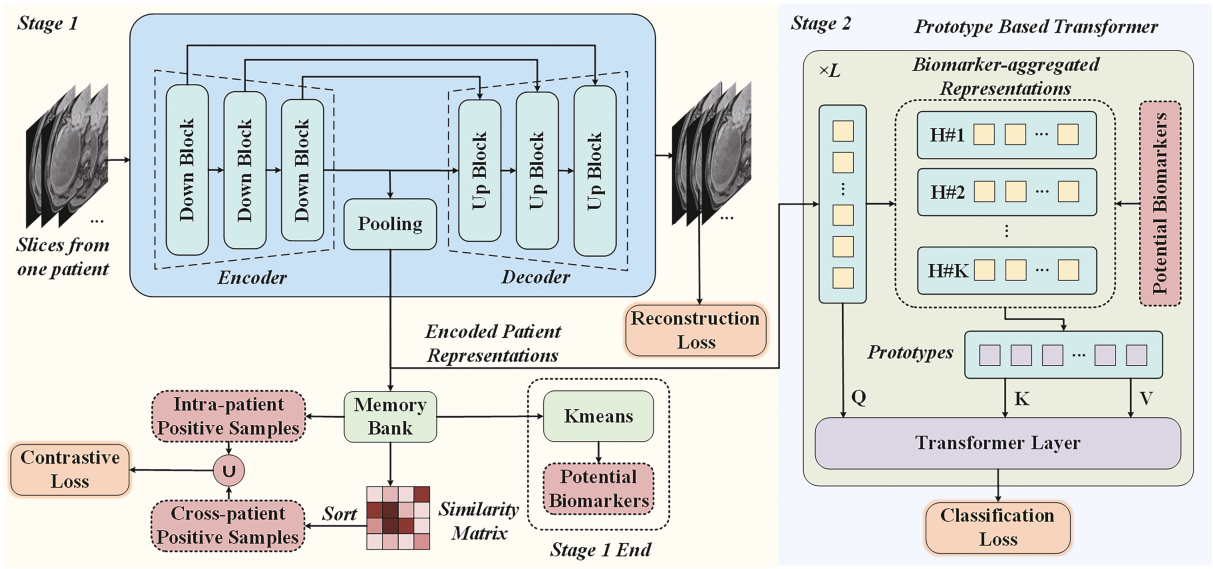

Figure 2: The overall architecture of the proposed IRCL network. Stage I employs a cascade encoder-decoder for latent representation extraction, regularized by image reconstruction. Stage II uses patient sequential data and a prototype biomarker attention mechanism to assess alignment between patient characteristics and identified biomarkers, enabling accurate prognostic prediction.

attributes of the extracted biomarkers and to capture the intrinsic distribution of the entire dataset, we devise a two-category contrastive learning method, which imposes constraints at both the slice level and the patient level. Upon completion of Stage I, all learned representations are aggregated and their cluster centers are defined as latent imaging biomarkers. Stage II takes patient sequential data as input. Here, we design a prototype biomarker attention mechanism to assess the alignment between individual patient characteristics and the candidate latent imaging biomarkers identified, allowing for accurate prediction of prognostic outcomes.

3.2 Stage I: Intra-Patient Level and inter-patient Level Contrastive Learning

This stage aims to learn interpretable latent representations from raw image data and extract latent imaging biomarkers via minimizing two kinds of losses: the image reconstruction loss $\mathcal{L}_c$ and the contrastive learning loss $\mathcal{L}_r$ as:

$$\mathcal{L} = \mathcal{L}_r + \mathcal{L}_c. \tag{1}$$

Here, the reconstruction term $\mathcal{L}_r$ ensures that all latent representations can be effectively mapped back to the original image space while constraining the features to retain sufficient image information. The contrastive term $\mathcal{L}_c$ aims to learn inter-patient generalized latent representations, thereby extracting common imaging patterns as biomarkers.

1) The Reconstruction Loss Define the dataset as $D = \{\cdots, \underbrace{I_k^r, I_{k+1}^r, \cdots, I_{k+m-1}^r}_{the\quad r_{th}\quad patient}, \cdots\}$, where $I_k^r \in R^{H\times W}$ is an image slice of the r-th patient, and each patient has the same number of slice images of m. These slices could originate from **different imaging modalities**. There are in total M patients. The downsampling path of the IRCL encoder is utilized to obtain intermediate image representations. Concurrently, the decoder path reconstructs generated images $\hat{I}_k$ from the intermediate representations, employing perceptual loss and L1 loss as loss functions to impose spatial regularization constraints. Let $f_i(I_k)$ and $f_i(\hat{I}_k)$ denote the features output from the i-th convolutional layer of a pre-trained Visual Geometry Group Network (VGG16) as f, and N be the total number of selected layers. The reconstruction loss can be defined as:

$$\mathcal{L}_r = \frac{1}{N}\sum_{i=1}^{N} \|f_i(I_k) - f_i(\hat{I}_k)\|_2^2 + \|I_k - \hat{I}_k\|_1. \tag{2}$$

2) The Contrastive Learning Loss To characterize the sample representations in the latent space and enhance the generalizability of the extracted biomarkers in Eq.2, inspired by (Kim et al. 2024), we propose a two-category contrastive learning strategy to capture two types of similarities: similarities among different slices of the same lesion (i.e., from one patient), and similarities among images from patients belonging to the same MVI category. Accordingly, we define two classes of positive samples as the intra-patient positive sample, and the inter-patient positive sample.

Intra-patient Positive Sample. Refers to images from different slices of the same lesion within the same patient. These images represent different imaging slices of the same HCC lesion, characterizing similarities among features belonging to the same patient. Denote $\mathcal{D}_k = \{I_k^r, I_{k+1}^r, \cdots, I_{k+m-1}^r\}$ the set of all image slices belonging to the lesion of image I_k^r, then any two images within $\mathcal{D}_k$ can be considered an intra-patient positive pair (e.g. for the k-th image, we have intra-patient positive sample pairs

as: $\{I_k^r, I_{k+1}^r\}$, $\{I_k^r, I_{k+2}^r\}$, etc.). These images are from the same patient. For a given MVI image, its corresponding intra-patient positive pairs are fixed.

Inter-patient Positive Sample. Defined as images belonging to the same MVI category. Given that image labels of MVI cannot be known beforehand in contrastive learning, without losing the generality, we assume that images from different patients with similar features belong to the same category. For image I_k^r of patient r, we compute its feature similarity with other patients and rank them to select the **s** most competitive features as inter-patient positive samples, denoted as $\mathcal{N}_k$. This implies that the inter-patient positive samples for each sample are not fixed; they aim to aggregate existing similarities among features from different patients to form more generalized feature representations.

Inter-patient Positive Sample Update and Loss Function. Given the dependency of inter-patient positive pairs on the similarities among existing sample embeddings, a gradual update strategy is designed for inter-patient positive samples as shown in **Alg. 1**. The Memory Bank comprises a Multilayer Perceptron (MLP) that maps all intermediate image representations into the embedding space. The mapped results are stored in an array $\mathbf{M}$ and maintained using a momentum update strategy. We first compute the similarity matrix $\mathbf{M}_{\text{sim}} \in R^{N\times N}$ for all samples in the embedding space (N is the total number of samples in the Memory Bank) as:

$$\mathbf{M}_{\text{sim}} = \text{softmax}(\mathbf{M}\mathbf{M}^\top/\tau), \tag{3}$$

where τ is the temperature parameter, which controls the smoothness of the similarity distribution. Each element $\mathbf{M}_{\text{sim}}(i,j)$ represents the similarity between sample i and sample j, typically calculated using cosine similarity. Building upon this, we calculate the entropy of each sample's similarity distribution. For each sample k, its normalized similarity distribution is $P_k = \mathbf{M}_{\text{sim}}(k,:)$. The entropy $H(P_k)$ of sample k's similarity distribution is defined as:

$$H(P_k) = -\sum_{j=1}^{N} P_k(j)\log P_k(j), \tag{4}$$

where $P_k(j)$ is the normalized similarity between sample k and sample j. We posit that lower entropy indicates a more concentrated similarity distribution represented by the embedding, signifying better learned feature representations. Therefore, the Memory Bank selects samples with lower entropy based on the proportion of the current training epoch to the total training epochs (of Stage I). Subsequently, among these selected samples, excluding the sample itself, the s most similar samples are identified as the inter-patient positive sample set for each sample. The contrastive learning loss $\mathcal{L}_c$ can be expressed as:

$$\mathcal{L}_c = -\log(\sum_{k=1}^{N}\sum_{j\in\mathcal{D}_k\cup\mathcal{N}_k} P_k(j)), \tag{5}$$

which include constrain to intra-patient sample from $\mathcal{N}_k$ and inter-patient sample from $\mathcal{D}_k$. In this way, $\mathcal{L}_c$ enables the encoder to learn effective features over slices level and patient level, thereby extracting more expressive image features and generating high-quality biomarkers.

Algorithm 1: inter-patient Positive Sample Update

Input: Memory Bank array: $\mathbf{M}$ (array of N embeddings), total training epochs in Stage I: E, temperature parameter: τ, number of most similar samples: s, update frequency: f
Output: inter-patient Positive Sample Set $\mathcal{N}$

Initialize $\mathcal{N} \leftarrow \emptyset$
for $epoch \leftarrow 1$ to E step f **do**
  Calculate $\mathbf{M}_{\text{sim}}$ as **Eq.3**
  $H \leftarrow \text{zeros}(N)$
  for $k \leftarrow 1$ to N **do**
    Update $H(k)$ as **Eq.4**
  end for
  $\mathbf{I}_{\text{sort}} \leftarrow \text{sort\_indices}(H, \textit{ascending})$
  for each $i \in \mathbf{I}_{\text{sort}}[1 \dots \lfloor N\cdot(epoch/E)\rfloor]$ **do**
    $P_i = \mathbf{M}_{\text{sim}}(i,:)$
    $P_i(i) = -\infty$
    $\mathcal{N}_i \leftarrow \text{sort\_indices}(P_i, \textit{descending})[1 \dots s]$
  end for
end for
return $\mathcal{N}$

Potential Biomarkers via Clustering. Upon completion of Stage I, all image embeddings are aggregated using the K-Means algorithm, based on their similarity measure in the latent space. This process yields K cluster centers, which represent common image features present in the dataset and are subsequently established as latent imaging biomarkers as $Bio = \{B_1, B_2, ..., B_K\}$.

3.3 Stage II: Biomarker-Prototype Based Prediction

As shown in Fig. 2, this stage aims to achieve a precise MVI prediction through similarity matching (alignment) between the input features and the learned latent imaging biomarkers. Notably, images from different modalities are treated as sequential inputs. Given an input of m images from the r-th patient, the trained IRCL encoder downsampling path extracts sequential image representations $H_r = [H_1, H_2, \dots, H_m] \in R^{m\times d}$. Based on the biomarkers derived from the trained contrastive learning module, the closest biomarker can be identified for each input feature by:

$$a_j = \underset{i}{\text{argmin}}\ B_i^\top H_j. \tag{6}$$

Here $a = [a_1, a_2, \cdots, a_m] \in R^m$ represents the closest biomarker of each image for the r-th patient. Specifically, within each training batch, we identify features belonging to the same biomarker and compute their centroid as the prototype for that biomarker. Subsequently, attention scores are calculated between each feature and its corresponding prototype. We introduce an indicator function $\mathbf{1}_{a_i=k}$, which signifies that it is effective only for features belonging to the biomarker k. The formula for calculating the prototype of the k-th biomarker is:

$$\bar{H}_k = \begin{cases} \frac{1}{N_k}\sum_{i=1}^{m}\mathbf{1}_{a_i=k}H_i, & if \quad N_k > 0 \\ 0, & otherwise. \end{cases} \tag{7}$$

Here, N_k is equal to $\sum_{i=1}^{m} \mathbf{1}_{a_i=k}$. The set of all biomarker prototypes is denoted as $\bar{H} = [\bar{H}_1, \bar{H}_2, \ldots, \bar{H}_m] \in R^{K \times d}$. Hence, with the network parameters W_Q, W_K and W_V, the prototype attention can be calculated as:

$$\text{ProtoAtten} = \text{Atten}(W_Q H, W_K \bar{H}, W_V \bar{H}), \quad (8)$$

where the attention function is defined as:

$$\text{Atten}(Q, K, V) = \text{softmax}(\frac{QK^\top}{\tau_a \sqrt{d}})V. \quad (9)$$

Here, τ_a represents the attention temperature between features and prototypes, controlling the softmax normalization operation within the attention mechanism. Finally, the output of the Transformer encoder is fed through a classification head, consisting of layer normalization and a linear layer, to yield the classification results.

3.4 Latent Biomarker Interpretation

To further elucidate the attribution regions of latent imaging biomarkers, we generate a learnable normalized mask $\mathbf{Q} \in [0,1]^{H \times W}$ for the input images of relevant patients, aiming to identify regions in the image that contribute most significantly to the target features. Given an input image I_k, its closest latent imaging biomarker is denoted as v_c. We multiply the mask with the image element-wise to obtain the masked image $I'_k = I_k \odot \mathbf{Q}$. The encoded embedding obtained from Stage I of the masked image is denoted as v_l. We optimize the following objective function using optimization methods such as gradient descent to iteratively guide the high-dimensional feature embedding v_l to closely approximate the imaging biomarker v_c:

$$\min_{\mathbf{Q}} \|v_l - v_c\|_1 + D_{KL}(\mathcal{N}(\mu_{\mathbf{Q}}, \sigma^2_{\mathbf{Q}}) \| \mathcal{N}(\mu_0, \sigma_0^2)). \quad (10)$$

Here, $\|\cdot\|_1$ denotes the L1 norm, measuring the distance between v_l and v_c. The KL divergence term D_{KL} quantifies the dissimilarity between the empirical Gaussian distribution of the mask $\mathbf{Q}$'s pixel values (with mean $\mu_{\mathbf{Q}}$ and variance $\sigma^2_{\mathbf{Q}}$) and $\mathcal{N}(\mu_0, \sigma_0^2)$. By minimizing this objective, the mask $\mathbf{Q}$ effectively identifies regions in the image that contribute most to the features, thereby visually showcasing the attribution regions of the imaging biomarkers.

3.5 Multi-modality IRCL

For multi-modality MRI images, our IRCL model adapts as follows: images from different modalities for each patient are concatenated along the channel dimension as input. In both intra-patient and inter-patient contrastive learning, the core computational methodology remains unchanged, with only the size of relevant sets potentially adjusted. Given the straightforward concatenation of input images, the specific modality from which a clustered latent imaging biomarker originates is dynamically determined by the algorithm during the learning process.

4 Experiments and Analysis

4.1 Experimental Setup

1) Datasets. **HCC-ZSH** (Sun et al. 2022) is an internal dataset collected by Zhongshan Hospital, comprising 121 patients, of whom 85 are with MVI. The dataset includes three modalities: T1, T1D and T1V, with each modality providing 3 slices. **HCC-WCH** (Pang et al. 2025) is a publicly available single-modality dataset collected by West China Hospital. Hepatobiliary Phase images were prioritized for initial processing and analysis. It consists of a total of 246 patients, including 110 with MVI and 136 without MVI. For both datasets, we used Regions of Interest from all images as the input for the model. To mitigate the impact of random factors on the experiments, all experiments were conducted using 5-fold cross-validation.

2) Experimental Setting. The model was trained on an L40 GPU, utilizing the PyTorch framework. We employed the Adam optimizer with an initial learning rate from $\{1 \times 10^{-4}, 5 \times 10^{-5}, 1 \times 10^{-5}\}$. Stage I training proceeded for 80 epochs, with a maximum of 120 epochs governed by an early stopping mechanism. To enhance model generalization, mixup data augmentation (Zhang et al. 2017) is applied to the training set, generating 30 times the number of mixed samples. The batch size is set to 128. In Stage I, the number of biomarkers is set to $K = 6$, and the number of inter-patient positives per sample is set to $s = 2$. The update frequency for inter-patient positives is set to 8 epochs. The momentum for the Memory Bank was 0.5, and the temperature parameter is set to $\tau = 0.07$. In Stage II, the attention temperature is set to $\tau_a = 0.5$.

3) Evaluation Metrics. Five commonly used evaluation metrics are employed to evaluate the classification performance, including accuracy (ACC), F1 score (F1), area under curve (AUC), precision (PRE) and recall (REC).

4.2 State-of-the-art Comparisons

To validate the proposed method, we compare it to following MVI prediction methods: ResNet (He et al. 2016), LA (Shi et al. 2020), BIF (Zheng, Yi, and Luo 2025), FG_SNet (Liu, Yi, and Luo 2024) and HA_CSL (Pang et al. 2025). We first conducted experiments on single-modality data, and the quantitative classification results of the proposed framework and comparison methods on HCC are presented in Tab. 1. On HCC-WCH single-modality dataset and the T1 modality of HCC-ZSH, it can be observed that our proposed method maintains relatively good levels in terms of ACC, F1 score, and AUC. However, on T1D and T1V modalities of the HCC-ZSH dataset, the performance improvement is relatively limited. We postulate that this is primarily due to the limited image pattern information contained within a single modality, coupled with the relatively short sequence length, which results in fewer decisive cues for the model. Subsequently, we also conducted experiments on the HCC-ZSH multi-modality dataset, and the experimental results demonstrate a significant performance improvement of our method as mentioned in Tab. 2. We attribute this enhancement to the effective complementarity of image patterns provided by different modalities, which in turn enables the model to learn a richer and more robust feature distribution in Stage I.

4.3 Ablation Study

This subsection evaluates the effectiveness of key modules in our framework: intra-patient/inter-patient positive sam-

Model	HCC-WCH			HCC-ZSH(T1)			HCC-ZSH(T1D)			HCC-ZSH(T1V)		
	ACC%	AUC%	F1%	ACC%	AUC%	F1%	ACC%	AUC%	F1%	ACC%	AUC%	F1%
ResNet	70.00	71.85	69.40	79.16	86.11	76.62	83.30	88.31	81.93	80.62	82.23	74.69
LA	73.33	82.59	71.84	76.66	76.94	75.46	76.66	80.83	74.98	79.16	78.61	78.76
BIF	–	–	–	85.51	87.39	88.29	86.67	88.45	**90.50**	84.35	87.27	**86.93**
FG_SNet	70.55	80.74	69.54	87.50	90.27	88.12	85.00	**90.55**	84.64	79.16	86.11	79.17
HA_CSL	70.55	80.37	70.01	81.66	90.55	81.17	79.16	86.11	79.17	75.00	78.05	74.30
IRCL	**80.08**	**86.61**	**80.59**	**91.11**	**90.86**	**88.93**	**88.06**	84.44	84.22	**89.04**	**93.53**	86.28

Table 1. Single-modality MVI classification performance comparisons of six methods on HCC-WCH and HCC-ZSH datasets. "-" denotes that the results are not available due to no code released.

Model	ACC%	AUC%	F1%	PRE%	REC%
ResNet	69.99	79.65	67.20	66.60	69.99
LA	73.33	79.09	67.22	73.01	73.33
BIF	89.75	87.70	89.75	**91.39**	89.75
FG_SNet	77.50	80.00	77.67	83.01	77.50
HA_CSL	73.33	77.22	73.39	75.22	73.33
IRCL	**92.78**	**91.26**	**90.39**	89.45	**92.78**

Table 2. Multi-modality MVI classification performance comparisons on HCC-ZSH dataset.

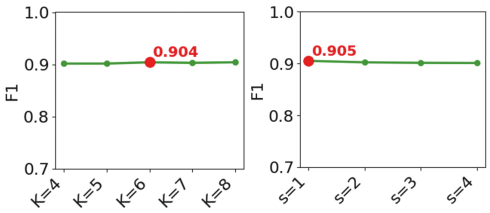

Figure 3: Sensitivity analysis of K and s on HCC-ZSH dataset.

ples, reconstruction loss, biomarker prototype attention and Mixup. Model performance upon module removal is shown in Tab. 3.

Model	ACC%	AUC%	F1%
w/o Intra-patient Sample	90.37	89.25	87.80
w/o inter-patient Sample	90.55	90.56	88.37
w/o Contrastive Loss $\mathcal{L}_c$	90.09	90.44	87.66
w/o Reconstruction Loss $\mathcal{L}_r$	89.35	90.65	87.80
w/o Prototype	92.17	90.98	90.10
w/o Mixup	79.81	89.19	73.16
IRCL	**92.78**	**91.26**	**90.39**

Table 3. Effectiveness of different components of our method on HCC-ZSH dataset. "w/o" denotes only this component excluded from our method.

Omitting Mixup significantly reduced accuracy, likely due to insufficient effective negative samples, limiting feature learning and generalization. Individually removing intra-patient and inter-patient positive sample constraints consistently degraded performance. We attribute this to intra-patient observations being crucial for capturing peritumoral morphological changes and reinforcing common patterns, while inter-patient observations enable generalized representations by associating features across patients. Furthermore, removing reconstruction loss also degraded performance, underscoring its pivotal role in Stage I by ensuring latent features retain sufficient image information, forming an MVI classification basis and acting as effective regularization for unsupervised training. Fig. 3 illustrates that hyperparamaters of s and K are not sensitive.

4.4 Interpretability of the Learned Biomarkers

Importance and Stability of Imaging Biomarkers. **Sec. 4.2** has demonstrated that the IRCL framework can predict MVI based on the excavated imaging biomarkers. However, *critical questions remain: Do these learned biomarkers hold statistical significance for MVI diagnosis across different patients? Which specific biomarkers determine the classification outcome for a given patient?* To address this, we fixed the patient features from the first-stage network, allowing us to treat the biomarkers associated with patient features as tabular feature vectors. We then applied the SHapley Additive exPlanations (SHAP) method (Lundberg and Lee 2017) to interpret the prediction ability of the model. Specifically, for any given prediction, SHAP calculates each biomarker's contribution to the prediction outcome (i.e., MVI positive and MVI negative). We statistically analyze the SHAP value distribution for each potential biomarker (i.e., Biomarker 0-5) as shown in Fig. 4(a) for HCC-ZSH multi-modal dataset. Meanwhile, given a biomarker and its SHAP histogram, we also compute the individual mean for the group of MVI-positive and MVI-negative patients; a mean apart from zero indicates the stronger discriminative capability. From Fig. 4(a), we can see that for the first three biomarkers, their SHAP values for negative patients tend towards zero, while significantly deviating from zero for positive patients, indicating specialization in MVI-positive features. Conversely, the latter three biomarkers show greater deviation from zero for negative patients than for positive patients, highlighting their focus on MVI-negative features. It further validates that multiple biomarkers jointly predict the MVI based on the multi-modality MRI data, which also reflects the prediction stability of each biomarker across diverse patients. Furthermore, in Fig. 4(b), an MVI positive

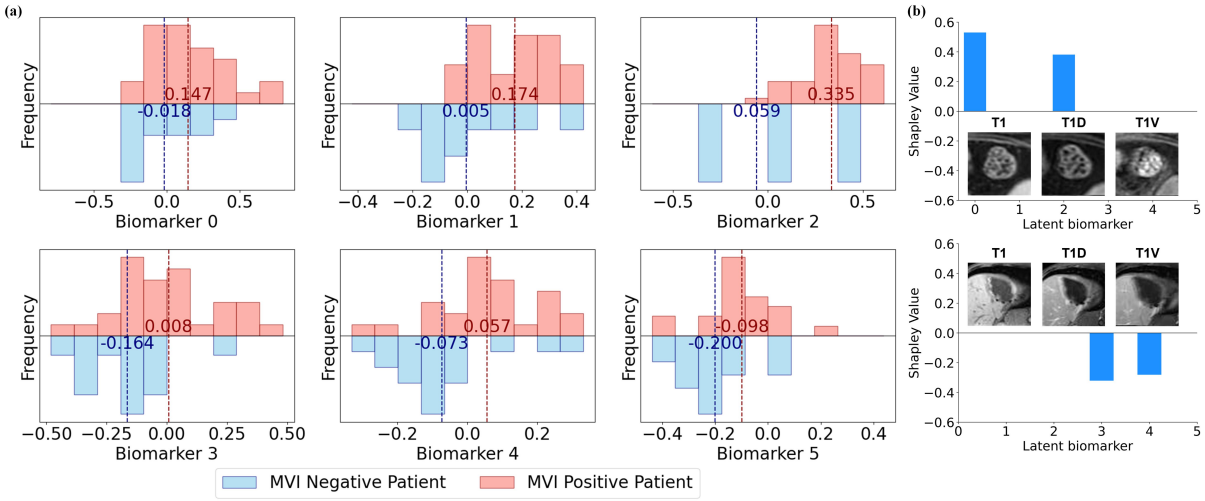

Figure 4: (a) Distribution of SHAP values for six biomarkers extracted by our IRCL method on HCC-ZSH dataset, differentiated by MVI-positive and MVI-negative patients. (b) Specific SHAP scores for individual biomarkers in MVI-positive and MVI-negative patients on HCC-ZSH dataset.

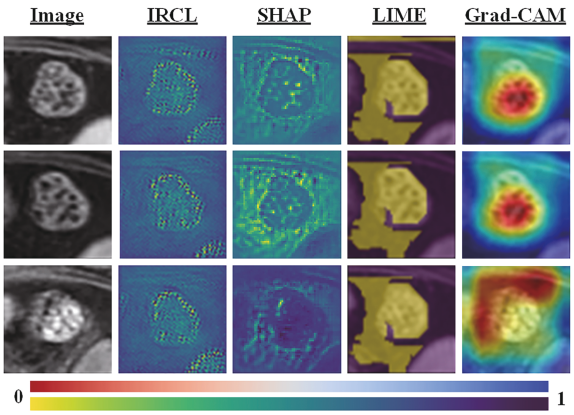

Figure 5: Visualization of the learned important regions by various methods on 3 MRI modality images (i.e., T1, T1D, T1V) for the patient (top) in Fig. 4(b) on HCC-ZSH dataset.

patient (the top) and a negative patient (the bottom) are list. The first three biomarkers contribute significantly for the MVI positive patient, while for the negative patient, the latter three show notable contributions, providing auxiliary diagnostic evidence. In summary, our excavated biomarkers demonstrate stability in diagnosing different patients and effectively indicate MVI status.

Imaging Biomarker Visualization. To elucidate the attribution regions of these biomarkers, we leverage the post-hoc analysis method described in **Sec. 3.4** to generate learnable attribution masks for images of MVI-positive patients. To validate the interpretability advantages of our method, we compare it with Grad-CAM (Selvaraju et al. 2017), SHAP (Lundberg and Lee 2017) and Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro, Singh, and Guestrin 2016) shown in Fig. 5. The interpretability results from Grad-CAM and LIME suffer from issues of excessively large attribution regions, posing difficulties for expert understanding. While the SHAP method can describe some important features, such as the liver capsule and liver parenchymal tissue, its consistency across different modalities is inferior compared to our biomarker-based explanations. By constraining the masks with imaging biomarkers as a reference, the learned masks can intuitively illustrate the regions of interest for the model, revealing the relationship between the regions of interest and structures. Specifically, an intact liver capsule typically suggests lower MVI risk, whereas an incomplete one may increase MVI likelihood; similarly, homogeneous liver tissue may indicate lower MVI tendency, while heterogeneous tissue often suggests higher aggressiveness (Wang et al. 2021). Our method provides such attribution information with clear clinical implications, thereby offering clinicians more comprehensible visualization results.

5 Conclusion

We propose IRCL, a novel interpretable framework for predicting MVI in HCC by learning latent imaging biomarkers. It employs a two-stage learning paradigm: Stage I uses dual-level contrastive learning to extract generalizable imaging biomarkers from MRI images, and Stage II predicts MVI based on the alignment between patient features and learned biomarkers. Additionally, a mask-based visual explanation method is designed to clearly delineate biomarkers in original images, improving model interpretability. Extensive experiments on two HCC datasets demonstrate IRCL's superior performance in both MVI prediction accuracy and interpretability compared to existing methods. This work offers valuable clinical support for HCC diagnosis and provides a foundation for future imaging biomarker discovery and explanation in other diseases.

Acknowledgements

This paper is supported by the General Program of the National Natural Science Foundation of China under Grant 62276189, Grants of Tongji University Medicine-X Interdisciplinary Research Initiative (2025-0554-YB-09 and 2025-0650-YB-17), and the Fundamental Research Funds for the Central Universities (2025-1-ZD-02).

References

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Kim, H.; Seo, C.; Cho, Y.; and Yoo, T. K. 2024. Patient-Level Contrastive Learning for Enhanced Biomarker Prediction in Retinal Imaging. In *Proceedings of the MICCAI Workshop on Data Engineering in Medical Imaging*, 125–133. Springer.

Liang, J.; Zhang, W.; Yang, J.; Wu, M.; Dai, Q.; Yin, H.; Xiao, Y.; and Kong, L. 2023. Deep learning supported discovery of biomarkers for clinical prognosis of liver cancer. *Nature Machine Intelligence*, 5(4): 408–420.

Liu, X.; Yi, Y.; and Luo, Y. 2024. A Cascade Multimodal Fine-Grained MRI Image Grading Network For Preoperative Microvascular Invasion In Hepatocellular Carcinoma. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 1–6. IEEE.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Mak, L.-Y.; Liu, K.; Chirapongsathorn, S.; Yew, K. C.; Tamaki, N.; Rajaram, R. B.; Panlilio, M. T.; Lui, R.; Lee, H. W.; Lai, J. C.-T.; et al. 2024. Liver diseases and hepatocellular carcinoma in the Asia-Pacific region: burden, trends, challenges and future directions. *Nature Reviews Gastroenterology & Hepatology*, 1–18.

O'connor, J. P.; Aboagye, E. O.; Adams, J. E.; Aerts, H. J.; Barrington, S. F.; Beer, A. J.; Boellaard, R.; Bohndiek, S. E.; Brady, M.; Brown, G.; et al. 2017. Imaging biomarker roadmap for cancer studies. *Nature reviews Clinical oncology*, 14(3): 169–186.

Pang, S.; Chen, Y.; Shi, X.; Wang, R.; Dai, M.; Zhu, X.; Song, B.; and Li, K. 2025. Interpretable 2.5 D network by hierarchical attention and consistency learning for 3D MRI classification. *Pattern Recognition*, 164: 111539.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Rotem, O.; Schwartz, T.; Maor, R.; Tauber, Y.; Shapiro, M. T.; Meseguer, M.; Gilboa, D.; Seidman, D. S.; and Zaritsky, A. 2024. Visual interpretability of image-based classification models by generative latent space disentanglement applied to in vitro fertilization. *Nature communications*, 15(1): 7390.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 618–626.

Shi, X.; Xing, F.; Xie, Y.; Zhang, Z.; Cui, L.; and Yang, L. 2020. Loss-based attention for deep multiple instance learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5742–5749.

Sun, B.-Y.; Gu, P.-Y.; Guan, R.-Y.; Zhou, C.; Lu, J.-W.; Yang, Z.-F.; Pan, C.; Zhou, P.-Y.; Zhu, Y.-P.; Li, J.-R.; et al. 2022. Deep-learning-based analysis of preoperative MRI predicts microvascular invasion and outcome in hepatocellular carcinoma. *World journal of surgical oncology*, 20(1): 189.

Wang, F.; Zhan, G.; Chen, Q.-q.; Xu, H.-y.; Cao, D.; Zhang, Y.-y.; Li, Y.-h.; Zhang, C.-j.; Jin, Y.; Ji, W.-b.; et al. 2024. Multitask deep learning for prediction of microvascular invasion and recurrence-free survival in hepatocellular carcinoma based on MRI images. *Liver International*, 44(6): 1351–1362.

Wang, W.; Guo, Y.; Zhong, J.; Wang, Q.; Wang, X.; Wei, H.; Li, J.; and Xiu, P. 2021. The clinical significance of microvascular invasion in the surgical planning and postoperative sequential treatment in hepatocellular carcinoma. *Scientific Reports*, 11(1): 2415.

Yao, L.-Q.; Li, C.; Diao, Y.-K.; Liang, L.; Jia, H.-D.; Tang, S.-C.; Zeng, Y.-Y.; Wu, H.; Wang, M.-D.; Gu, L.-H.; et al. 2023. Grading severity of microscopic vascular invasion was independently associated with recurrence and survival following hepatectomy for solitary hepatocellular carcinoma. *Hepatobiliary Surgery and Nutrition*, 13(1): 16.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, S.; Shi, T.; Jiang, Y.; Zhang, X.; Lei, J.; Feng, Z.; and Song, M. 2023. A Loopback Network for Explainable Microvascular Invasion Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7443–7453.

Zhang, Y.; Liu, H.; Zhu, L.; Chong, H.; Fu, H.; Yu, L.; Li, P.; Qin, J.; Feng, D. D.; and Wang, L. 2024. Modality-aware Distillation Network for Microvascular Invasion Prediction of Hepatocellar Carcinoma from MRI Images. *IEEE Transactions on Biomedical Engineering*.

Zhao, A.; Shahin, A. H.; Zhou, Y.; Gudmundsson, E.; Szmul, A.; Mogulkoc, N.; Van Beek, F.; Brereton, C. J.; Van Es, H. W.; Pontoppidan, K.; et al. 2022. Prognostic imaging biomarker discovery in survival analysis for idiopathic pulmonary fibrosis. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 223–233. Springer.

Zheng, P.; Yi, Y.; and Luo, Y. 2025. BIF: A Biosignature Identification Framework for Model-agnostic Interpretation of MVI Diagnosis Models in HCC. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. IEEE.

Zheng, T.; Zhu, Y.; Jiang, H.; Yang, C.; Ye, Y.; Bashir, M. R.; Li, C.; Long, L.; Luo, S.; Song, B.; et al. 2025. MRI-Based Topology Deep Learning Model for Noninvasive Prediction of Microvascular Invasion and Assisting Prognostic Stratification in HCC. *Liver International*, 45(3): e16205.

Zhou, G.; Zhou, Y.; Xu, X.; Zhang, J.; Xu, C.; Xu, P.; and Zhu, F. 2024. MRI-based radiomics signature: a potential imaging biomarker for prediction of microvascular invasion in combined hepatocellular-cholangiocarcinoma. *Abdominal Radiology*, 49(1): 49–59.