



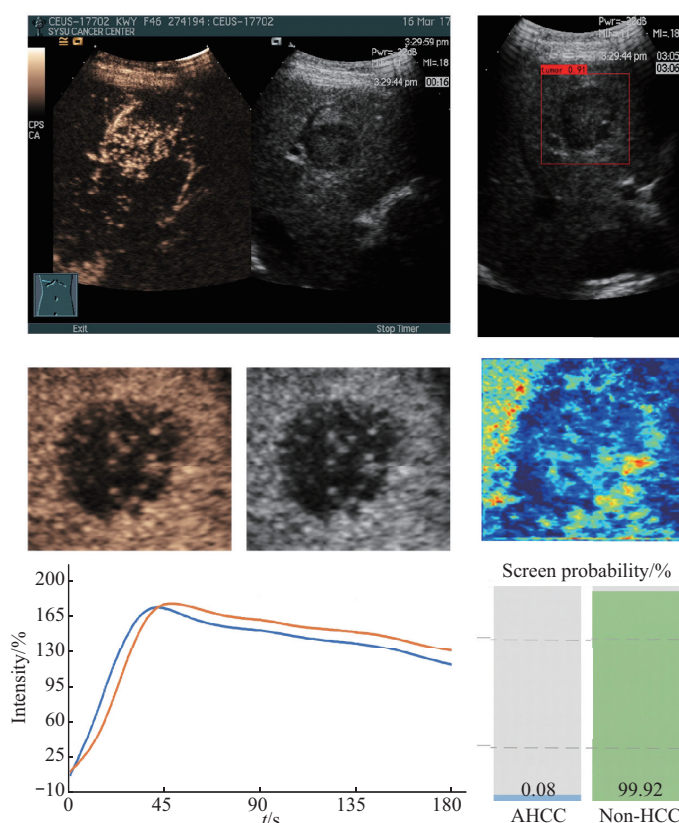
An Attention-weighted Tri-modal Ultrasound Network (TUS-Net) for Screening of Atypical Hepatocellular Carcinoma From LR-M Liver Nodules*

ZHANG He-Chong^{2)**}, HUANG Liang-Hui^{2)**}, WANG Xue-Hua^{2)***}, JIANG Shang-Lin²⁾,
 CHEN Ying-Ying²⁾, ZENG Ya-Guang²⁾, ZHENG Wei^{1)***}

⁽¹⁾*Collaborative Innovation Center of Cancer Medicine, State Key Laboratory of Oncology in South China, Department of Ultrasound, Sun Yat-sen University Cancer Center, Guangzhou 510060, China;*

⁽²⁾*School of Physics and Optoelectronic Engineering, Foshan University, Foshan 528225, China)*

Graphical abstract



* This work was supported by a grant from The National Natural Science Foundation of China (82202179).

** These authors contributed equally to this work.

*** Corresponding author.

WANG Xue-Hua. Tel: 86-757-82716895, E-mail: xhwang10000@163.com

ZHENG Wei. Tel: 86-20-87343212, E-mail: zhengwei@sysucc.org.cn

Received: January 24, 2026 Accepted: April 15, 2026

Abstract Objective Discriminating atypical hepatocellular carcinoma (HCC) from other malignancies in liver nodules classified as Liver Imaging Reporting and Data System category M (LR-M) remains a significant diagnostic challenge on conventional ultrasound examination. The LR-M category, originally intended to capture non-HCC malignancies, paradoxically contains up to 63% of atypical HCCs that deviate from classic enhancement patterns, leading to potential misdiagnosis and suboptimal treatment planning. While deep learning has shown promise in HCC diagnosis, most existing models rely exclusively on single-modality ultrasound, overlooking the diagnostic benefits of integrating complementary information from multiple imaging sources. To address this gap, we propose a novel attention-weighted tri-modal ultrasound network (TUS-Net) that integrates contrast-enhanced ultrasound (CEUS), B-mode ultrasound (BUS), and time-intensity curves (TICs) to improve diagnostic accuracy for these clinically challenging lesions. **Methods** Our framework incorporates a three-dimensional convolutional neural network (C3D) backbone to extract spatiotemporal features from CEUS videos, capturing dynamic vascular patterns critical for lesion characterization. To effectively fuse complementary modalities, we introduce a dual-channel feature fusion module (DCFFM) that adaptively combines features from CEUS and BUS through channel-wise attention mechanisms, allowing the model to dynamically weigh the contribution of each modality based on diagnostic relevance. Additionally, we propose a temporal intensity feature fusion module (TIFFM) that leverages quantitative hemodynamic information from TICs to guide the model's attention toward diagnostically critical temporal phases, such as arterial wash-in and portal venous washout. The model is further enhanced by automated lesion localization using YOLOX and class activation mapping for interpretability, ensuring that predictions align with clinically meaningful imaging features. **Results** Evaluated on a tri-modal ultrasound dataset comprising 161 patients with pathologically confirmed LR-M nodules (131 atypical HCC and 30 non-HCC malignancies), our model achieved an accuracy of 86.83%, a sensitivity of 92.50%, a specificity of 75.50%, and an *AUC* of 89.32% in screening atypical HCC. Compared to single-modality baselines, TUS-Net demonstrated superior specificity, a clinically critical metric given the higher risk associated with misclassifying non-HCC malignancies. Ablation studies confirmed the contribution of each module, with the full model outperforming both standard C3D and 3D ResNet backbones integrated with attention mechanisms. A reader study involving junior and senior radiologists further validated the clinical utility of AI assistance, showing consistent improvements in specificity and inter-reader consistency, particularly for less experienced clinicians. **Conclusion** These results surpass existing benchmark models and demonstrate the potential of our approach to enhance diagnostic precision in clinically specific cases. By intelligently fusing multi-modal ultrasound data with attention-guided mechanisms, TUS-Net offers a reliable and interpretable tool that holds promise for improving the non-invasive diagnosis of atypical HCC in challenging LR-M liver nodules.

Key words deep learning, attention mechanism, hepatocellular carcinoma, ultrasound

DOI: 10.3724/j.pibb.2026.0038

CSTR: 32369.14.pibb.20260038

Primary liver cancer, predominantly comprising hepatocellular carcinoma (HCC) and intrahepatic cholangiocarcinoma (ICC), presents a significant global health challenge^[1]. While histopathological analysis remains the diagnostic gold standard for focal liver lesions (FLLs)^[2-3], its invasive nature has motivated the development of reliable non-invasive alternatives. Contrast-enhanced ultrasound (CEUS) has emerged as a pivotal modality in this context, providing real-time visualization of tumor vascular architecture and hemodynamics^[4-5].

The CEUS liver imaging reporting and data system (LI-RADS) establishes standardized criteria for interpreting liver nodules in at-risk patients, with the LR-M category designated for lesions that are likely malignant but exhibit features atypical for HCC^[6]. This classification presents a particular

diagnostic dilemma: while demonstrating high sensitivity for non-HCC malignancies, the LR-M category contains a substantial proportion (up to 63%) of atypical HCC cases^[7-8]. These atypical HCCs deviate from the classic enhancement pattern of arterial hyperenhancement followed by washout, instead showing iso-enhancement or persistent hyperenhancement. This overlap in imaging features leads to misclassification into LR-M and consequently diminishes the positive predictive value of this category^[8]. The clinical implications are significant, as treatment strategies for HCC and non-HCC malignancies diverge considerably^[9], making accurate discrimination within LR-M nodules a pressing clinical priority.

The inherent heterogeneity of liver tumors further compounds this diagnostic challenge^[10].

Different ultrasound techniques provide complementary information: B-mode ultrasound (BUS) excels in depicting morphological features and tumor boundaries^[11]; CEUS captures dynamic vascular patterns^[11-12], and time-intensity curves (TIC) derived from CEUS offer quantitative hemodynamic profiles^[13]. While radiologists routinely synthesize information from these modalities, the manual integration of high-dimensional data is cognitively demanding and susceptible to inter-observer variability^[14].

Deep learning approaches, particularly convolutional neural networks (CNNs), present a promising solution for this complex task^[15-16]. Current CNN-based methodologies for liver lesion analysis can be categorized into several approaches. Single-modality methods employing 2D CNNs for BUS^[15] and computed tomography (CT)^[17-18], or 3D architectures like three-dimensional convolutional neural network (C3D)^[19] for dynamic magnetic resonance imaging (MRI)/CT data^[20], have demonstrated substantial diagnostic capability. However, their reliance on a single data source limits their ability to capture the complementary information utilized in clinical practice. Multi-modal fusion strategies have emerged to address this limitation, with techniques integrating multi-phase magnetic resonance (MR)^[21] or combining CT with pathology data^[22] showing improved performance^[23]. In ultrasound, radiomics models incorporating BUS, CEUS and TIC features have demonstrated particular value^[24-25]. Nevertheless, these approaches often lack adaptive mechanisms to dynamically weight the diagnostic contribution of each modality. Concurrently, attention mechanisms have enhanced feature representation in medical imaging through modules like squeeze-and-excitation (SE)^[26] and convolutional block attention module (CBAM)^[27], though their application in ultrasound remains limited, particularly for incorporating clinical priors such as temporal enhancement patterns.

To address these limitations, we propose a novel attention-weighted tri-modal ultrasound network (TUS-Net). Our framework introduces two key innovations: a dual-channel feature fusion module (DCFFM) that adaptively integrates spatiotemporal features from CEUS with structural features from BUS, and a temporal intensity feature fusion module (TIFFM) that leverages TIC dynamics to guide the

model's attention to diagnostically critical temporal phases. The architecture is complemented by automated lesion localization using YOLOX^[28] and class activation mapping (CAM)^[29] for decision interpretability. By developing this targeted solution, we aim to enhance diagnostic precision for challenging LR-M nodules and support more informed clinical decision-making.

1 Patients and data

1.1 Study subjects

This study was conducted utilizing a retrospectively compiled dataset from patients who underwent both conventional ultrasound and CEUS examinations at Sun Yat-sen University Cancer Hospital between September 2014 and June 2019. Initially, 373 patients were screened according to predefined inclusion criteria. After rigorous evaluation, the final cohort consisted of 161 liver nodules with arterial phase hyperenhancement (APHE) that were classified as LR-M according to CEUS LI-RADS criteria. This specific population was selected due to the high prevalence of atypical HCC within this category.

The study protocol received approval from the Human Research Ethics Committee of Sun Yat-sen University Cancer Center and complied with the principles of the Declaration of Helsinki (GYX2020-017). All patient data were anonymized prior to analysis. The inclusion criteria were: (1) availability of complete CEUS examination with diagnostic quality; (2) age $\geq$ 18 years; (3) no history of liver biopsy or prior treatment before ultrasound examination; (4) presence of LR-M classified nodules exhibiting APHE based on LI-RADS guidelines; and (5) definitive diagnosis confirmed by needle biopsy or surgical pathology.

1.2 Acquisition of B-mode ultrasound and contrast-enhanced ultrasound data

BUS examinations were performed using an Acuson Sequoia 512 system (Siemens Medical Solutions, Mountain View, USA) equipped with a 4C1 convex array transducer. Standardized scanning protocols were implemented to document lesion characteristics including quantity, dimensions, spatial location, echogenicity, and background liver parenchyma.

CEUS was conducted using low mechanical

index ($MI < 0.2$) imaging with a dynamic range of 80 dB. After establishing baseline BUS characteristics, a 2.0 ml bolus of SonoVue (Bracco Imaging, Milan, Italy) was administered *via* antecubital venous access, immediately followed by a 5-ml saline flush. Continuous cine-loop acquisition was initiated upon contrast agent injection and maintained for 80 s to capture the complete arterial, portal venous, and late vascular phases. Consistent machine parameters were preserved throughout the acquisition. Subsequent intermittent scanning was

performed over a 5-min period to characterize lesion enhancement patterns and document washout features.

2 Methods

Figure 1 depicts a schematic of our proposed methodology, segmented into three main phases: localizing the liver lesion region (Figure 1a), creating a tri-modal ultrasound dataset (Figure 1b), and developing deep learning models for classification (Figure 1c–f).

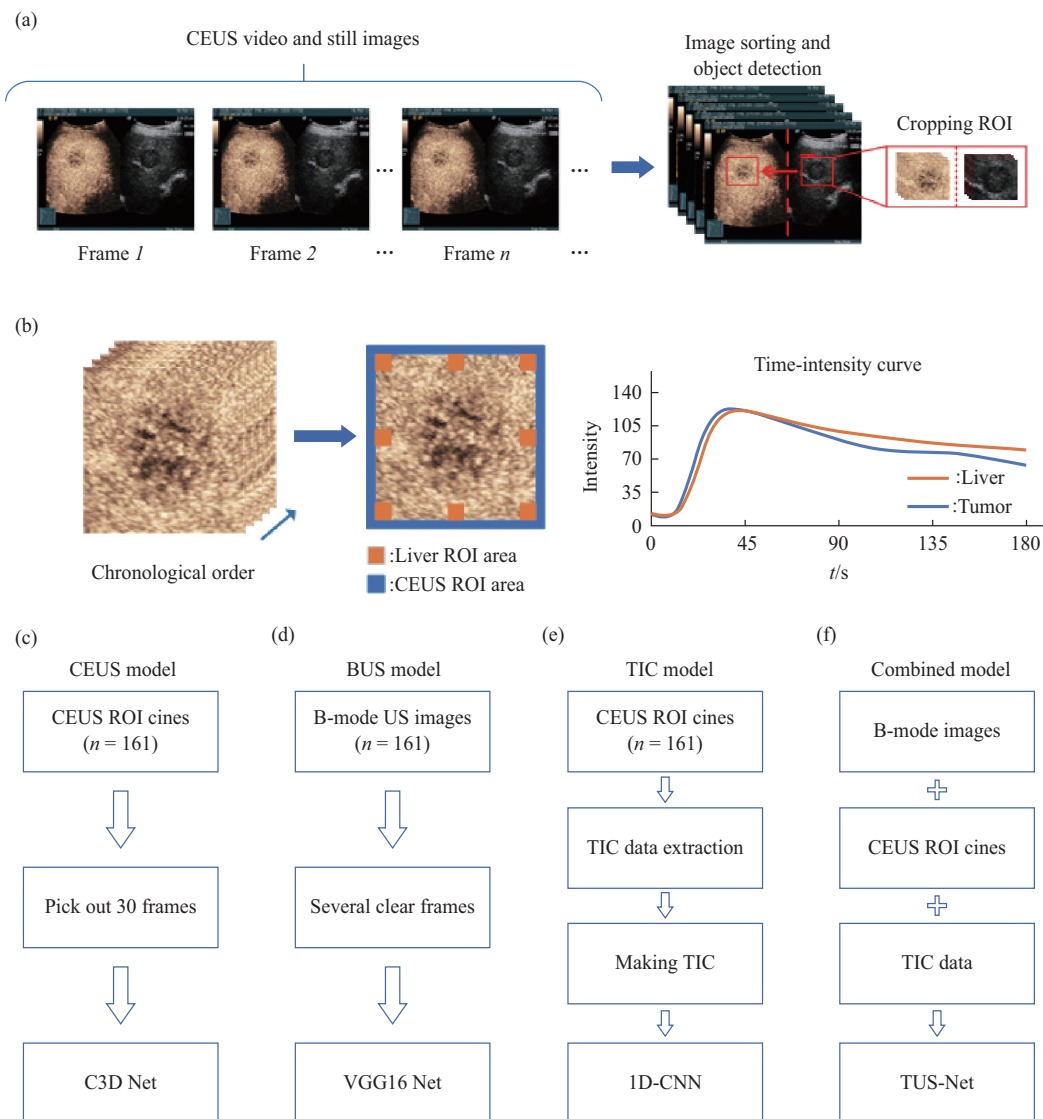


Fig. 1 Methodological workflow

(a) Data preprocessing and region-of-interest (ROI) extraction from that integrates contrast-enhanced ultrasound (CEUS) and B-mode ultrasound (BUS) sequences. (b) Construction of the tri-modal ultrasound dataset. (c) A three-dimensional convolutional neural network (C3D) is implemented to extract spatiotemporal features from CEUS videos. (d) A VGG16-based model is used to extract features from BUS images^[30]. (e) A 1D-CNN is employed to analyze time-intensity curves (TICs)^[30]. (f) The proposed Tri-modal Ultrasound Network (TUS-Net) integrates features from all three modalities for final classification.

2.1 Automated lesion localization using YOLOX

Since extraneous image margins introduced unwanted interference during model training, we developed an automated lesion localization pipeline on BUS using the YOLOX architecture^[28], leveraging BUS's superior tumor boundary delineation^[31].

The region of interest (ROI) extraction protocol proceeded as follows (Figure 2a). After preprocessing BUS images with noise reduction and margin removal, an experienced radiologist annotated rectangular bounding boxes around liver lesions. These annotations were guided by simultaneous review of CEUS cine sequences to ensure accurate lesion identification. The annotation protocol followed three key principles: (1) standardized labeling criteria across all cases; (2) inclusion of approximately 25% tumor area within each ROI; and (3) central positioning of lesions within ROIs.

This process generated 4 830 annotated BUS images from 161 patients. The dataset was randomly partitioned into training, validation, and testing subsets following a 6 : 2 : 2 ratio. The trained YOLOX network was then deployed to automatically localize lesions in all remaining BUS images, eliminating manual annotation requirements for the full dataset. The corresponding ROI in the CEUS sequence was spatially aligned and transferred using the coordinates obtained from the BUS image.

2.2 Attention-weighted tri-modal ultrasound network (TUS-Net) for the fusion of multimodal feature

2.2.1 Tri-modal ultrasound dataset

The tri-modal ultrasound dataset integrates three complementary data sources: dynamic CEUS videos, BUS images, and quantitative TICs. The data preprocessing and construction of this dataset were conducted according to the methodology established in our prior work^[32-33].

For CEUS video processing, all ROI images were converted to grayscale and standardized to 128×128 pixels. Video sequences were constructed starting from the onset of tumor enhancement, with temporal sampling at 2-s intervals generating approximately 25 frames. These were combined with static images to form standardized 30-frame CEUS videos. To address class imbalance (atypical HCC: 75%–85% vs. non-HCC: 15%–25%), comprehensive data augmentation was implemented. Horizontal flipping was applied

universally, while non-HCC cases received additional augmentation through 1-s interval resampling. This strategy effectively doubled HCC cases and increased non-HCC cases sixfold in the training set.

BUS ROIs were similarly resized to 128×128 pixels. To maintain class balance during training, we systematically selected 2 atypical HCC and 6 non-HCC images per case through random sampling.

TIC extraction employed a dual-region approach: normal liver parenchyma intensity was averaged from eight peripheral boxes within the CEUS ROI (Figure 2b), while tumor intensity was computed from the central region excluding these boxes. The resulting curves were refined using Gaussian filtering and polynomial fitting techniques to minimize noise while preserving essential hemodynamic patterns.

2.2.2 Attention-weighted tri-modal ultrasound network (TUS-Net) architecture

Figure 2 outlines the structure of the TUS-Net model, consisting of four key components: the backbone C3D for processing CEUS video (CEUS model), the modified VGG16 network for BUS images (BUS model), the DCCFM, and the TIFFM.

The C3D backbone comprises three sets of 3D convolution and pooling layers, mainly used to extract features from CEUS videos. The BUS model is a customized version of the VGG16 model^[30]. It includes six convolutional layers and three pooling layers, each with a kernel and pooling size of 3×3.

The TIFFM is designed to inject temporal attention into the C3D backbone. Formally, let $F_{c3d} \in R^{T \times H \times W \times C}$ represent the spatiotemporal feature maps extracted by the C3D backbone, where T denotes temporal length, $H \times W$ spatial dimensions, and C channel depth. Simultaneously, the TIC of tumor ROI are divided into 10 equal time gradients to reflect the speed of intensity changes, which is encoded as $V_{tic} \in R^{D_t}$, where D_t represents the dimensionality of TIC features.

The temporal attention mechanism operates through the following transformation:

$$W_t \in R^{T \times 2C} a_{c3d} = GAP_{time}(F_{c3d}) \in R^C \quad (1)$$

$$a_{tic} = W_c \cdot V_{tic} \in R^C \quad (2)$$

$$A_t = \sigma \cdot (W_t \cdot (a_{c3d} \oplus a_{tic})) \in R^T \quad (3)$$

Where $GAP_{time}(\cdot)$ denotes global average pooling along the temporal dimension, a_{c3d} and a_{tic} are features of the temporal dimension from CEUS and TIC, respectively. $W_c \in R^{C \times D_t}$ and $W_t \in R^{T \times 2C}$ are learnable

projection matrices, $\oplus$ indicates concatenation, and σ represents the sigmoid activation function.

The computed temporal attention weights A_t are then applied to recalibrate the original features:

$$F_{\text{c3d}}^{\text{att}} = A_t \otimes F_{\text{c3d}} \quad (4)$$

where $\otimes$ denotes element-wise multiplication with broadcasting along spatial and channel dimensions. This operation effectively emphasizes crucial temporal phases in CEUS.

The DCFFM enables adaptive integration of features from CEUS and BUS modalities. Let

$F_{\text{ceus}} \in R^{H' \times W' \times C'}$ and $F_{\text{bus}} \in R^{H' \times W' \times C'}$ represent the reshaped feature maps from CEUS and BUS streams respectively, after dimensional alignment.

The feature fusion process begins with concatenation:

$$F_{\text{cat}} = [F_{\text{ceus}} \oplus F_{\text{bus}}] \in R^{H' \times W' \times 2C'} \quad (5)$$

Channel attention weights are derived through squeeze-and-excitation operations:

$$z = \text{GAP}_{\text{spatial}}(F_{\text{cat}}) \in R^{2C'} \quad (6)$$

$$A_c = \delta \cdot (W_2 \cdot \delta(W_1 \cdot z)) \in R^{2C'} \quad (7)$$

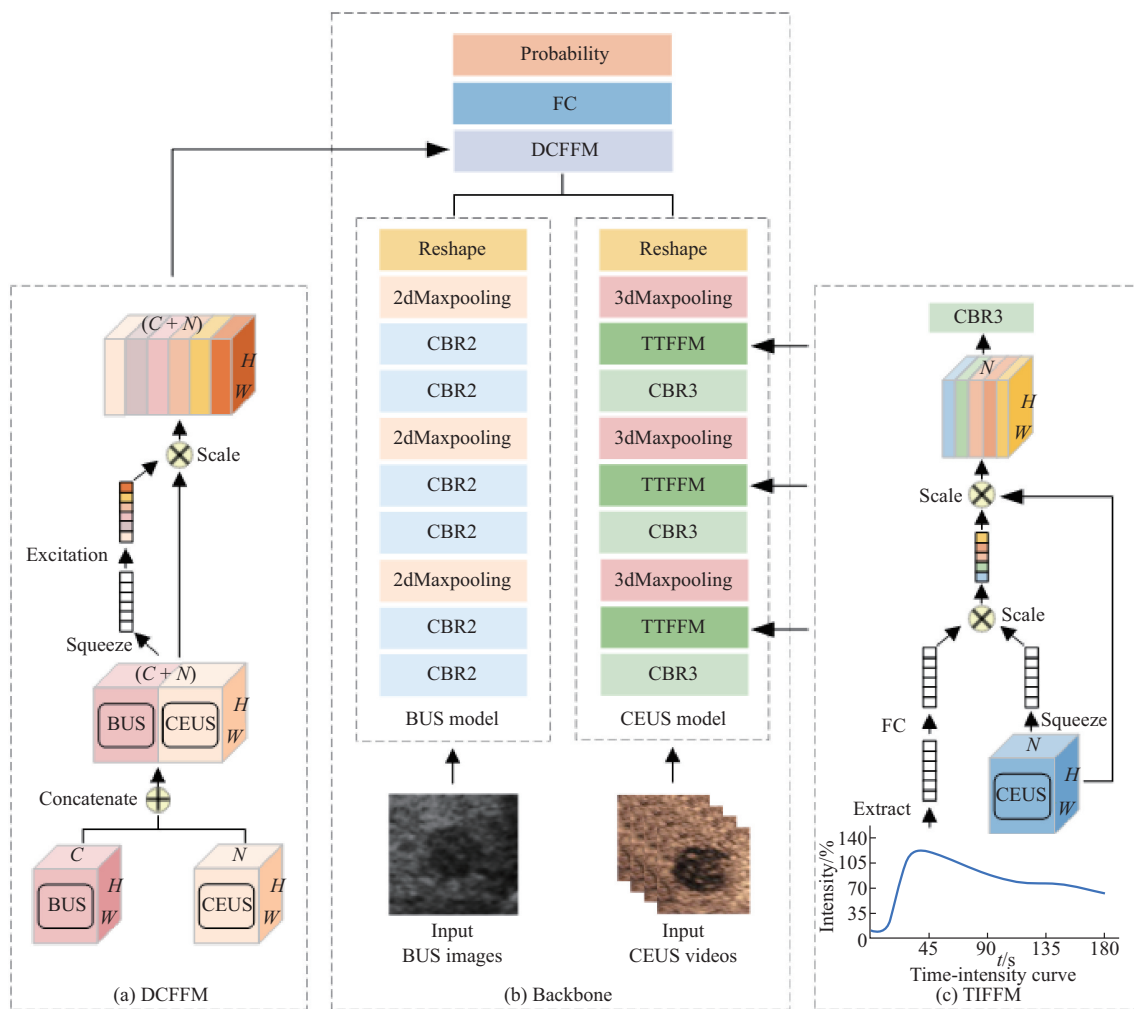


Fig. 2 Architecture of the attention-weighted tri-modal ultrasound network

The proposed framework integrates four key components: (1) a C3D backbone employing three sequential 3D convolutional and pooling blocks for spatiotemporal feature extraction from CEUS videos; (2) a customized VGG16 network with six convolutional and three pooling layers (all 3×3 kernels) for BUS feature extraction; (3) the Temporal Intensity Feature Fusion Module (TIFFM) that computes temporal attention weights by integrating time-intensity curve (TIC) features with temporally-pooled C3D features; (4) the Dual-Channel Feature Fusion Module (DCFFM) that adaptively fuses modality-specific features through channel-wise attention mechanisms. The model ultimately produces a 512-dimensional vector for final classification. Notation: CBR2/CBR3 means 2D/3D convolutional layers with batch normalization and ReLU activation; FC is fully connected layer; TIFFM is temporal intensity feature fusion module; DCFFM is dual-channel feature fusion module. All convolutional and pooling layers utilize 3×3 kernels.

where GAP_{spatial} denotes global average pooling across spatial dimensions, $W_1 \in R^{\frac{2C'}{r} \cdot 2C'}$ and $W_2 \in R^{\frac{2C'}{r} \cdot 2C'}$ are learnable parameters with reduction ratio r , and δ represents the ReLU activation function.

The attention-weighted features are computed as:

$$F_{\text{weighted}} = A_c \otimes F_{\text{cat}} \in R^{H' \times W' \times 2C'} \quad (8)$$

Finally, the module produces the fused representation through dimensionality reduction:

$$F_{\text{fusion}} = W_{\text{fusion}} \cdot \text{Flatten}(F_{\text{weighted}}) \in R^{512} \quad (9)$$

where $W_{\text{fusion}} \in R^{512 \times (H' \times W' \times 2C')}$ projects the weighted features to a 512-dimensional vector for final classification.

This dual-attention architecture enables both temporal focus on enhancement patterns and adaptive weighting of modality-specific contributions, effectively mimicking the clinical decision-making process of expert radiologists.

3 Experiments and results

3.1 Experimental setting

All models in this study were constructed using the PyTorch framework and optimized for performance leveraging an NVIDIA GeForce RTX 2080Ti 11 GB GPU. For fine-tuning, the Adam optimizer was employed, with an initial learning rate set at $1e-5$, adjusted through cosine annealing. To

assess the localization accuracy of the YOLOX model, average precision (AP) and accuracy metrics were employed^[31]. For the evaluation of the classification models, a comprehensive set of metrics was utilized, including accuracy (ACC), precision (PPV), negative predictive value (NPV), sensitivity (SE), specificity (SP), F1-score, receiver operating characteristic (ROC) and area under the ROC Curve (AUC). The screening dataset consisted of tri-modal US data of 161 patients, including 131 atypical HCC patients and 30 non-HCC patients. We formed 20 testing sets to evaluate model performance by randomly selecting 20 atypical HCC and 10 Non-HCC from the screening dataset. Other 131 cases that did not overlap with the testing set were divided into training and validation sets in a ratio of 8 : 2. Given the class imbalance between atypical HCC and non-HCC cases, data augmentation techniques such as random horizontal flipping, random Gaussian noise injection, and histogram equalization were applied during training. Additionally, to effectively address the class imbalance, binary weighted cross-entropy loss was used, setting a weight coefficient of 1 : 3 for atypical HCC and non-HCC categories, respectively.

3.2 Localization results of liver lesions on BUS

Figure 3 show six instances illustrating the localization outcomes of YOLOX (depicted in

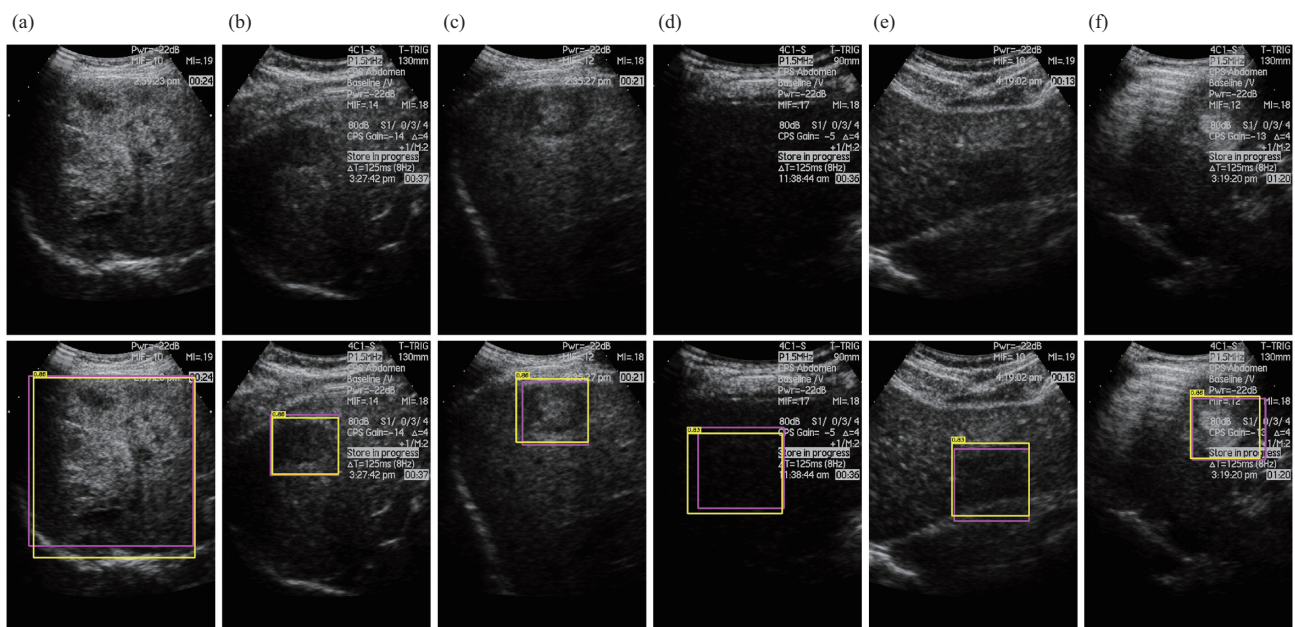


Fig. 3 Localization results of liver lesions using YOLOX and radiologist annotation

The upper row shows the BUS images, and the lower row shows the localization results of YOLOX and radiologist annotation results. The yellow bounding box ROI was the YOLOX output results, and the purple bounding box ROI was the radiologist manual delineation of liver lesions. (a-c) shows the BUS images with clearer tumor location. (d-f) shows the BUS image of tumor difficult for human eyes to recognize.

yellow) and radiologists (represented in purple) for liver lesions. These examples underscore a remarkable alignment between YOLOX predictions and radiologist annotations, especially in challenging scenarios where human identification was difficult (Figure 3d,e), or obstructed by text from instruments (Figure 3f). In a test set of 966 BUS images, YOLOX achieved an accuracy of 96.07% and an AP value of 97.70%, demonstrating its proficiency in accurately identifying and localizing liver tumors and surrounding tissues.

3.3 Screening results of atypical HCC from LR–M liver lesions on tri–modal ultrasound data

Table 1 displays the performance of six classification models applied to the tri-modal ultrasound data. The single mode TIC model, BUS

model, and CEUS model achieved average accuracies of 75.00% (95% confidence intervals (CI), 70.43%–79.57%), 81.83% (95% CI, 78.61%–85.05%), and 84.00% (95% CI, 79.72%–88.28%), respectively. The CEUS model outperformed the others for the CEUS data contains key hemodynamic characteristics. Incorporating the TIFFM and DCFFM into the CEUS model enhanced accuracy to 86.33% and 85.67%, respectively. Furthermore, the TUS-Net achieved the highest accuracy of 86.83%, precision of 88.77%, specificity of 75.50%, and F1-score of 90.36%, with moderate NPV of 84.52% and sensitivity of 92.50%. The ROC curves in Figure 4a underscore the superior AUC of the TUS-Net compared to others, affirming the pronounced contribution of tri-modal ultrasound fusion in enhancing the classification capabilities.

Table 1 Performance comparison of six classification models on the tri–modal ultrasound test set (95% confidence intervals are shown in parentheses)

Model	Accuracy/%	Precision/%	NPV/%	Sensitivity/%	Specificity/%	F1-score/%
TIC model	75.00 (70.43–79.57)	82.73 (78.77–86.70)	63.68 (55.74–71.63)	80.25 (74.35–86.15)	64.50 (55.46–73.54)	80.70 (76.77–84.62)
BUS model	81.83 (78.61–85.05)	86.09 (83.14–89.03)	74.25 (68.82–79.68)	87.25 (84.08–90.42)	71.00 (64.23–77.77)	86.48 (84.09–88.87)
CEUS model	84.00 (79.72–88.28)	84.64 (80.88–88.40)	82.48 (73.57–91.39)	93.75 (91.13–96.37)	64.50 (54.60–74.40)	88.79 (85.98–91.61)
CEUS+ DCFFM	85.67 (82.27–89.07)	85.93 (82.72–89.15)	85.75 (79.92–91.59)	94.50 (92.51–96.49)	68.00 (59.88–76.12)	89.89 (87.62–92.17)
CEUS+ TIFFM	86.33 (83.02–89.65)	86.72 (83.63–89.82)	87.30 (80.91–93.70)	94.50 (91.67–97.33)	70.00 (62.26–77.74)	90.26 (87.96–92.56)
TUS-Net	86.83 (83.65–90.01)	88.77 (85.67–91.87)	84.52 (77.94–91.10)	92.50 (89.32–95.68)	75.50 (68.14–82.86)	90.36 (88.11–92.61)

The bold numbers in the table indicate the highest values.

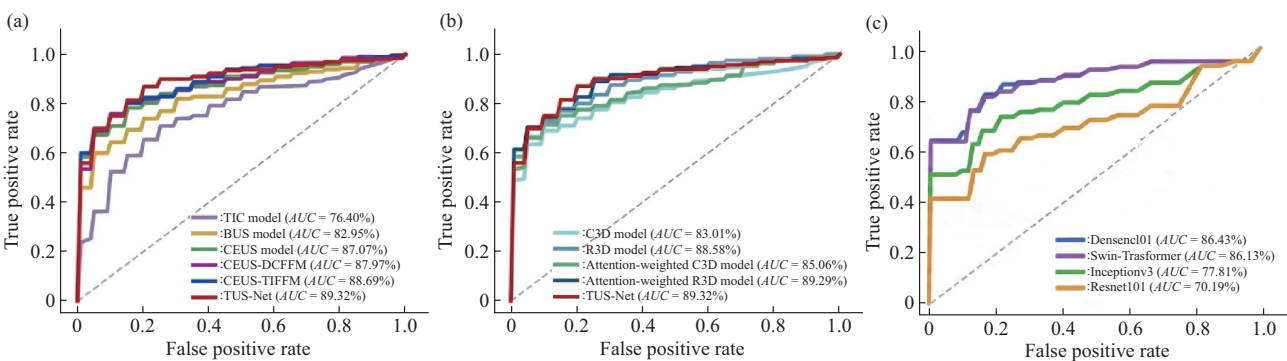


Fig. 4 ROC curves for classification performance

(a) Comparison of the six models in Table 1; (b) backbone and attention-weighted fusion comparisons in Table 2; (c) comparison with mainstream models in Table 3.

Additional studies were undertaken to validate the efficacy of the proposed attention-weighted mechanism for tri-modal ultrasound fusion. Comparisons were made using previous C3D^[19] and 3D ResNet (R3D) models^[34], as detailed in Table 2. These comparisons involved integrating the TIFFM and DCFFM into both the C3D and R3D models. The outcomes demonstrate that the integrated models

improved in terms of accuracy (by 2.67% and 2.34%), specificity (by 6.5% and 4.5%), and AUC by 2.05% and 0.71%, respectively. These outcomes confirm the effectiveness of the attention-weighted mechanism in enhancing various backbone models. Notably, our model achieved the highest AUC value among all tested models, as shown in Figure 4b.

Table 2 Performance of the attention-weighted fusion method across different backbone networks (95% confidence intervals for quantitative data are shown in parentheses)

Model	Accuracy/%	Precision/%	NPV/%	Sensitivity/%	Specificity/%	F1-score/%
C3D model ^[19]	82.00	82.76	81.97	92.25	61.50	87.14
	(77.43–86.57)	(79.72–85.80)	(72.99–90.95)	(87.86–96.64)	(54.50–68.50)	(83.70–90.57)
R3D model ^[34]	83.83	84.49	83.35	93.50	64.50	88.60
	(79.71–87.95)	(81.02–87.96)	(74.79–91.92)	(90.45–96.55)	(55.46–73.54)	(85.82–91.38)
Attention-weighted C3D model	84.67	85.54	83.71	93.00	68.00	89.00
	(80.93–88.40)	(82.47–88.62)	(77.11–90.30)	(89.83–96.17)	(60.94–75.06)	(86.28–91.73)
Attention-weighted R3D model	86.17	86.37	87.59	94.75	69.00	90.18
	(82.20–90.13)	(82.92–89.83)	(80.52–94.67)	(91.49–98.01)	(60.29–77.71)	(87.38–92.98)
TUS-Net	86.83	88.77	84.52	92.50	75.50	90.36
	(83.65–90.01)	(85.67–91.87)	(77.94–91.10)	(89.32–95.68)	(68.14–82.86)	(88.11–92.61)

The bold numbers in the table indicate the highest values.

To further evaluate the performance of the proposed TUS-Net against established mainstream models, we performed an auxiliary comparative study using sparsely sampled static frames. For each case, three representative CEUS frames (the 2nd, 15th, and 30th) were selected to construct a frame-based dataset. Multiple architectures including Swin-Transformer^[35], DenseNet201^[36], Inceptionv3^[37], and ResNet101^[38], were then trained and tested on this dataset. As shown

in Table 3, these baselines attained average accuracies between 69.28% and 79.17%, while TUS-Net retained a distinct advantage with an accuracy of 86.83%. This superiority is further confirmed by the ROC curves in Figure 4c. These findings demonstrate that utilizing complete information from tri-modal ultrasound, offers substantially stronger discriminative capability compared to frame-based, unimodal approaches.

Table 3 Auxiliary comparison with mainstream models using sparsely sampled frames (95% confidence intervals for quantitative data are shown in parentheses)

Model	Accuracy/%	Precision/%	NPV/%	Sensitivity/%	Specificity/%	F1-score/%
Swin-Transformer ³⁵	79.17	85.91	67.42	82.25	73.00	84.02
	(74.44–82.22)	(83.64–88.89)	(60.00–75.00)	(76.67–88.33)	(70.00–80.00)	(80.00–86.89)
Resnet101 ³⁶	69.28	73.18	58.83	85.42	37.00	78.61
	(60.00–75.56)	(68.75–76.92)	(38.46–78.57)	(71.67–96.67)	(13.33–53.33)	(70.97–83.82)
Densenet201 ³⁷	78.89	73.77	86.53	89.17	58.33	84.91
	(74.44–87.78)	(62.50–84.62)	(82.35–87.50)	(80.00–96.67)	(36.67–80.00)	(81.36–90.91)
Inceptionv3 ³⁸	74.83	77.45	67.99	88.00	48.50	82.30
	(70.00–81.11)	(73.61–81.16)	(56.52–84.62)	(80.00–96.67)	(36.67–60.00)	(78.74–86.82)

4 Model interpretability and clinical application

4.1 Visualizing the decision of TUS-Net

To illustrate how the TUS-Net distinguishes and categorizes features from tri-modal ultrasound data, we generated heatmaps using class activation mapping (CAM)^[29] for both atypical HCC (upper panel) and non-HCC (lower panel) cases. These heatmaps, alongside the original CEUS images for comparisons, are displayed in Figure 5. The colored bars in these heatmaps indicate the significance level of the contributing features, with the highest value marked as 1 and the lowest as 0. This visualization highlights the pivotal role of these features in the model's classification process. These CAM images provide preliminary insights into distinct patterns:

(1) The initial heatmaps revealed that atypical

HCC cases exhibited scattered low-intensity heating values within the tumor region, whereas non-HCC cases showed high-intensity heating values concentrated along the rim region.

(2) In the middle and late phases of the heatmaps, the highlighted regions predominantly encompassed the tumor borders for atypical HCC cases and the tumor interiors for non-HCC cases.

(3) This distinction in the highlighted regions within the heatmaps between atypical HCC and non-HCC cases is noteworthy, with non-HCC cases generally showing larger highlighted regions compared to atypical HCC cases.

These patterns correspond closely to established CEUS diagnostic criteria for differentiating atypical HCC from non-HCC malignancies^[39], demonstrating that the model has successfully captured the essential radiological features.

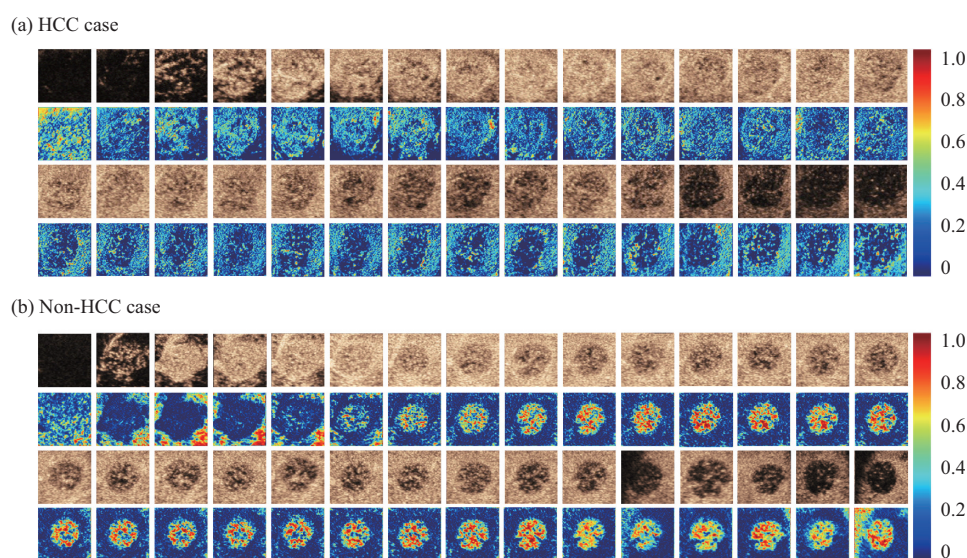


Fig. 5 Examples of CAM heatmaps generated by attention-weighted TUS-Net for atypical HCC and non-HCC

4.2 Clinical evaluation of AI-assisted diagnosis

To assess the clinical utility of TUS-Net, we conducted a reader study with junior and senior radiologists comparing unassisted and AI-assisted performance. As summarized in Table 4, AI-assistance consistently reduced false-positive findings, improving specificity for both readers across cohorts. In the training cohort, the senior radiologist achieved gains in both specificity (86.67%→96.67%) and sensitivity (97.71%→99.24%), while the junior

radiologist showed a slight sensitivity increase with stable specificity. In the validation cohort, specificity improved for both readers, though sensitivity remained unchanged for the junior reader and decreased slightly for the senior reader, indicating occasional trade-offs in difficult cases.

Figure 6 presents representative cases of AI-assisted interpretation, illustrating how model outputs prompt re-evaluation of ambiguous LR-M lesions. While not infallible, the AI-generated heatmaps and

scores supported more confident final decisions under clinician oversight, with particularly noticeable benefit for the junior radiologist. Together, these findings demonstrate that TUS-Net can improve

diagnostic consistency and reduce inter-reader variability, with its impact modulated by reader experience and case complexity.

Table 4 Performance improvement of junior and senior radiologists with AI assistance

	Readers	True negative	True positive	Specificity	Sensitivity
Training cohort	Junior	16→16	121→122 ↑	53.33%→53.33%	92.37%→93.13% ↑
	Senior	26→29 ↑	128→130 ↑	86.67%→96.67% ↑	97.71%→99.24% ↑
Validation cohort	Junior	17→18 ↑	11→11	80.95%→85.71% ↑	78.57%→78.57%
	Senior	17→20 ↑	13→12	80.95%→95.24% ↑	92.86%→85.71%

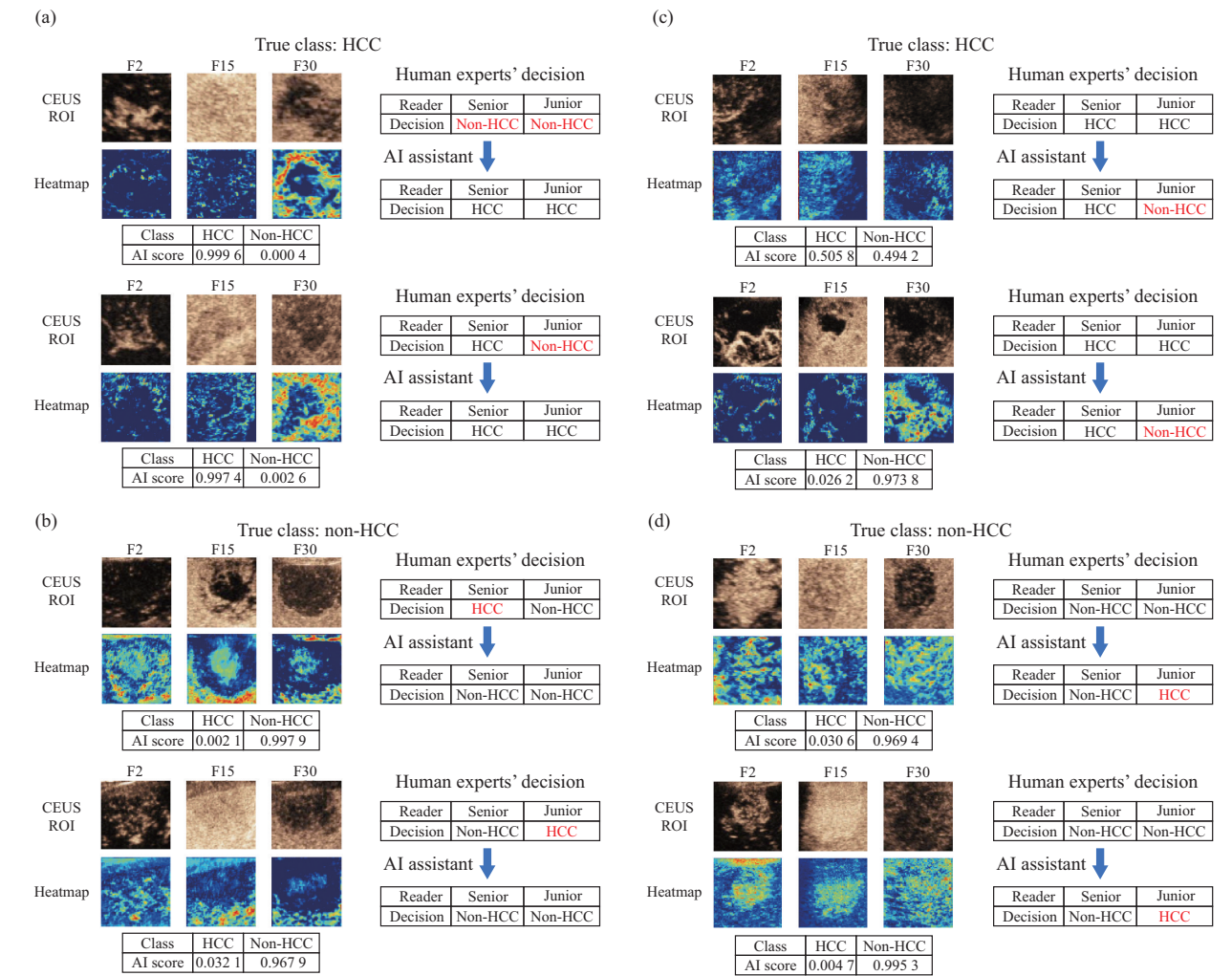


Fig. 6 Impact of TUS-Net on diagnostic decision-making in LR-M cases

(a) False-negative correction: the model prompted reevaluation of a missed HCC *via* highlighted intratumoral enhancement. (b) False-positive correction: identification of rim-like enhancement led to reversal of a false-positive call. (c) Borderline confusion: ambiguous AI output resulted in incorrect diagnosis revision by a junior reader. (d) Misinterpretation: correct AI suggestion was misinterpreted, causing erroneous decision shift by a junior reader. These cases demonstrate that AI assistance should augment, not replace clinical judgment. Red text: incorrect decisions; black text: correct decisions.

5 Discussion

This study proposes TUS-Net, an attention-weighted tri-modal ultrasound network for the precise identification of atypical HCC within diagnostically ambiguous LR-M liver nodules. The method combines automated lesion localization via YOLOX, and two dedicated attention modules, the TIFFM and the DCFM, to effectively integrate complementary information from CEUS videos, BUS images, and TICs.

Experimental results validate the efficacy of the proposed framework. TUS-Net achieved an accuracy of 86.83% and a specificity of 75.50% in distinguishing atypical HCC from non-HCC malignancies within the LR-M category, surpassing existing benchmarks. While TUS-Net achieves slightly lower NPV and sensitivity than certain comparative models, this trade-off is clinically justified: false positives carry greater risk, as non-HCC malignancies require radical surgery whereas atypical HCC can be managed with ablation^[9]. By prioritizing specificity to minimize high-risk errors, TUS-Net accepts a modest sensitivity reduction. Moreover, visualization through class activation mapping confirmed that the model's decision-making aligns with established radiological patterns, thereby enhancing its clinical interpretability and trustworthiness.

The clinical relevance of this work is underscored by its utility as a decision-support tool, especially for less-experienced radiologists. By providing quantitative predictions and visually salient heatmaps, TUS-Net guides diagnostic focus and improves interpretability. The notable performance gap compared to state-of-the-art models, not originally designed for ultrasound analysis, further highlights the importance of developing modality-specific architectures.

We acknowledge that this study is retrospective and relies on a single-center dataset of limited size, which may affect model generalizability. Future work will aim to expand the dataset through multi-center collaborations and to explore the integration of additional imaging modalities, such as MRI and CT, toward building a more comprehensive diagnostic

system.

6 Conclusion

TUS-Net advances the non-invasive diagnosis of atypical HCC by intelligently fusing tri-modal ultrasound data through a tailored attention-weighted architecture. It offers a reliable, interpretable computational tool that represents a meaningful step toward more accurate and accessible diagnosis of challenging LR-M liver nodules.

Data Availability The associated data for this paper (DOI: 10.57760/sciencedb.33446/CSTR: 31253.11.sciencedb.33446) is freely accessible in the Science Data Bank (ScienceDB) database at the following link: <https://www.scidb.cn/detail?dataSetId=82329e41b77943cb836fd7797a2d6374>.

References

- [1] Sung H, Ferlay J, Siegel R L, *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, 2021, **71**(3): 209-249
- [2] Altekruse S F, McGlynn K A, Dickie L A, *et al.* Hepatocellular carcinoma confirmation, treatment, and survival in surveillance, epidemiology, and end results registries, 1992-2008. *Hepatology*, 2012, **55**(2): 476-482
- [3] di Tommaso L, Spadaccini M, Donadon M, *et al.* Role of liver biopsy in hepatocellular carcinoma. *World J Gastroenterol*, 2019, **25**(40): 6041-6052
- [4] Jang H J, Kim T K, Wilson S R. Small nodules (1-2 Cm) in liver cirrhosis: characterization with contrast-enhanced ultrasound. *Eur J Radiol*, 2009, **72**(3): 418-424
- [5] Hanna R F, Miloussev V Z, Tang A, *et al.* Comparative 13-year meta-analysis of the sensitivity and positive predictive value of ultrasound, CT, and MRI for detecting hepatocellular carcinoma. *Abdom Radiol (NY)*, 2016, **41**(1): 71-90
- [6] Kono Y, Lyshchik A, Cosgrove D, *et al.* Contrast enhanced ultrasound (CEUS) liver imaging reporting and data system (LI-RADS®): the official version by the American college of radiology (ACR). *Ultraschall Med*, 2017, **38**(1): 85-86
- [7] Dietrich C F, Nolsøe C P, Barr R G, *et al.* Guidelines and good clinical practice recommendations for contrast-enhanced ultrasound (CEUS) in the liver – update 2020 WFUMB in cooperation with EFSUMB, AFSUMB, AIUM, and FLAUS. *Ultrasound Med Biol*, 2020, **46**(10): 2579-2604
- [8] Zheng W, Li Q, Zou X B, *et al.* Evaluation of contrast-enhanced US LI-RADS version 2017: application on 2020 liver nodules in patients with hepatitis B infection. *Radiology*, 2020, **294**(2): 299-307

- [9] Jang H J, Kim T K, Burns P N, *et al.* Enhancement patterns of hepatocellular carcinoma at contrast-enhanced US: comparison with histologic differentiation. *Radiology*, 2007, **244**(3): 898-906
- [10] Losic B, Craig A J, Villacorta-Martin C, *et al.* Intratumoral heterogeneity and clonal evolution in liver cancer. *Nat Commun*, 2020, **11**(1): 291
- [11] Zhou J, Sun H C, Wang Z, *et al.* Guidelines for diagnosis and treatment of primary liver cancer in China (2017 edition). *Liver Cancer*, 2018, **7**(3): 235-260
- [12] Salvatore V, Gianstefani A, Negrini G, *et al.* Imaging diagnosis of hepatocellular carcinoma: recent advances of contrast-enhanced ultrasonography with SonoVue®. *Liver Cancer*, 2016, **5**(1): 55-66
- [13] Huang-Wei C, Bleuzen A, Bourlier P, *et al.* Differential diagnosis of focal nodular hyperplasia with quantitative parametric analysis in contrast-enhanced sonography. *Invest Radiol*, 2006, **41**(3): 363-368
- [14] Wan P, Xue H, Liu C, *et al.* Transport-based anatomical-functional metric learning for liver tumor recognition using dual-view dynamic CEUS imaging. *IEEE Trans Biomed Eng*, 2023, **70**(3): 1012-1023
- [15] Lee J H, Joo I, Kang T W, *et al.* Deep learning with ultrasonography: automated classification of liver fibrosis using a deep convolutional neural network. *Eur Radiol*, 2020, **30**(2): 1264-1273
- [16] Nishida N, Yamakawa M, Shiina T, *et al.* Current status and perspectives for computer-aided ultrasonic diagnosis of liver lesions using deep learning technology. *Hepatol Int*, 2019, **13**(4): 416-421
- [17] Midya A, Chakraborty J, Srouji R, *et al.* Computerized diagnosis of liver tumors from CT scans using a deep neural network approach. *IEEE J Biomed Health Inform*, 2023, **27**(5): 2456-2464
- [18] Othman E, Mahmoud M, Dhahri H, *et al.* Automatic detection of liver cancer using hybrid pre-trained models. *Sensors (Basel)*, 2022, **22**(14): 5429
- [19] Tran D, Bourdev L, Fergus R, *et al.* Learning spatiotemporal features with 3D convolutional networks//IEEE. 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015: 4489-4497
- [20] Zhou Q, Zhou Z, Chen C, *et al.* Grading of hepatocellular carcinoma using 3D SE-DenseNet in dynamic enhanced MR images. *Comput Biol Med*, 2019, **107**: 47-57
- [21] Li S, Xie Y, Wang G, *et al.* Adaptive multimodal fusion with attention guided deep supervision net for grading hepatocellular carcinoma. *IEEE J Biomed Health Inform*, 2022, **26**(8): 4123-4131
- [22] Khan R A, Fu M, Burbridge B, *et al.* A multi-modal deep neural network for multi-class liver cancer diagnosis. *Neural Netw*, 2023, **165**: 553-561
- [23] Wang Z, Wu J, Li C, *et al.* Diagnosis of osteosarcoma based on multimodal microscopic imaging and deep learning. *J Innov Opt Health Sci*, 2025, **18**(2): 2343001
- [24] Liu D, Liu F, Xie X, *et al.* Accurate prediction of responses to transarterial chemoembolization for patients with hepatocellular carcinoma by using artificial intelligence in contrast-enhanced ultrasound. *Eur Radiol*, 2020, **30**(4): 2365-2376
- [25] Huang Q, Pan F, Li W, *et al.* Differential diagnosis of atypical hepatocellular carcinoma in contrast-enhanced ultrasound using spatio-temporal diagnostic semantics. *IEEE J Biomed Health Inform*, 2020, **24**(10): 2860-2869
- [26] Hu J, Shen L, Sun G. Squeeze-and-excitation networks//IEEE. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 7132-7141
- [27] Woo S, Park J, Lee J Y, *et al.* CBAM: convolutional block attention module//Springer. Computer Vision-ECCV 2018. Munich Germany: Springer, 2018: 3-19
- [28] Li Y, Zhang Z, Sun J, *et al.* Tumor detection based on deep mutual learning in automated breast ultrasound. *Multimedia Tools and Applications*, 2024, **83**(26): 68421-68439
- [29] Selvaraju R R, Cogswell M, Das A, *et al.* Grad-CAM: visual explanations from deep networks *via* gradient-based localization. *Int J Comput Vis*, 2020, **128**(2): 336-359
- [30] Wang M, Li D, Tang L. A lung image classification method: a classifier constructed by combining improved vgg16 and gradient boosting decision tree. *J Mech Med Biol*, 2021, **21**(4): 2150042
- [31] Guan C, He B, Zhang H, *et al.* Label-free *in-vivo* classification and tracking of red blood cells and platelets using Dynamic-YOLOv4 network. *J Innov Opt Health Sci*, 2024, **17**(5): 2450009
- [32] 李睿, 许祥丛, 林静怡, 等. 基于卷积神经网络的超声影像肝癌自动分类. *生物化学与生物物理进展*, 2023, **50**(3): 668-675
- [33] Li R, Xu X C, Lin J Y, *et al.* *Prog Biochem Biophys*, 2023, **50**(3): 668-675
- [33] 陈楹楹, 江尚霖, 黄良汇, 等. 基于超声造影图像亮度变化曲线的肝细胞癌预测. *生物化学与生物物理进展*, 2025, **52**(8): 2163-2172
- [34] Chen Y Y, Jiang S L, Huang L H, *et al.* *Prog Biochem Biophys*, 2025, **52**(8): 2163-2172
- [34] Tran D, Wang H, Torresani L, *et al.* A closer look at spatiotemporal convolutions for action recognition//IEEE. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 6450-6459
- [35] Liu Z, Lin Y, Cao Y, *et al.* Swin transformer: hierarchical vision transformer using shifted windows//IEEE. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 9992-10002
- [36] Huang G, Liu Z, Van Der Maaten L, *et al.* Densely connected convolutional networks//IEEE. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 2261-2269
- [37] Szegedy C, Vanhoucke V, Ioffe S, *et al.* Rethinking the inception architecture for computer vision//IEEE. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 2818-2826
- [38] He K, Zhang X, Ren S, *et al.* Deep residual learning for image recognition.//IEEE. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 770-778
- [39] Wildner D, Bernatik T, Greis C, *et al.* CEUS in hepatocellular carcinoma and intrahepatic cholangiocellular carcinoma in 320 patients - early or late washout matters: a subanalysis of the DEGUM multicenter trial. *Ultraschall Med*, 2015, **36**(2): 132-139

注意力加权三模态超声网络 (TUS-Net) 用于 筛查LR-M型肝结节中非典型肝细胞癌*

张贺崇^{2)**} 黄良汇^{2)**} 王雪花^{2)***} 江尚霖²⁾ 陈楹楹²⁾ 曾亚光²⁾ 郑 玮^{1)***}

⁽¹⁾ 中山大学肿瘤防治中心超声科, 华南肿瘤学国家重点实验室, 肿瘤医学省部共建协同创新中心, 广州 510060;

⁽²⁾ 佛山大学物理与光电工程学院, 佛山 528225)

摘要 **目的** 准确区分LR-M型肝结节中的非典型肝细胞癌(HCC)与其他恶性肿瘤仍是超声诊断的难点。现有深度学习模型多依赖单一模态, 未充分挖掘多模态信息的互补价值。为此, 本文提出一种注意力加权三模态超声网络(TUS-Net), 融合多模态超声信息以提升对非典型HCC的鉴别能力。**方法** 该模型融合超声造影(CEUS)、B型超声(BUS)及时间-强度曲线(TIC)三种模态。采用C3D主干提取CEUS视频的时空特征; 通过双通道特征融合模块(DCCFM)自适应整合CEUS与BUS特征; 利用时序强度特征融合模块(TIFFM)引导模型关注TIC所反映的关键时相变化。此外, 模型结合YOLOX实现病灶自动定位, 并采用类激活映射提升决策可解释性。**结果** 在包含161例患者的三模态超声数据集上, TUS-Net筛查非典型HCC的准确率达86.83%, 敏感性为92.50%, 特异性为75.50%, *AUC*为89.32%。消融实验验证了各模块的有效性, 研究表明AI辅助可提升临床医生(尤其是低年资医师)的诊断特异性。**结论** 该模型表现优于现有基准方法, 表明多模态融合结合注意力机制的策略可有效提升疑难肝结节的诊断准确率, 为临床提供辅助决策支持。

关键词 深度学习, 注意力机制, 肝细胞癌, 超声

中图分类号 R445.1

DOI: 10.3724/j.pibb.2026.0038

CSTR: 32369.14.pibb.20260038

* 国家自然科学基金(82202179)资助项目。

** 并列第一作者。

*** 通讯联系人。

王雪花 Tel: 0757-82716895, E-mail: xhwang10000@163.com

郑玮 Tel: 020-87343212, E-mail: zhengwei@sysucc.org.cn

收稿日期: 2026-01-24, 接受日期: 2026-04-15