KG-MMIL: A Knowledge-Guided Multi-modal Multi-Instance Learning Framework for Interpretable Renal Tumor Ultrasound Diagnosis

Wei Ni[1], Tao Yang[1✉] and Kunlei Tan[2]

[1]School of Computer Science and Artificial Intelligence, Hunan University of Technology, Hunan 412007, China
[2]Zhuzhou Central Hospital, Zhuzhou 412007, Hunan, China
1768746087@qq.com

Abstract. Current methods for diagnosing renal tumors via ultrasound primarily rely on learning image features; however, the models' decision-making process lacks adequate interpretability, exhibiting typical "black-box" characteristics that limit their clinical application. To address this issue, this paper proposes a knowledge-guided multi-modal multi-instance learning fusion model (KG-MMIL) for enhancing the interpretability of renal tumor ultrasound diagnosis.First, a knowledge-guided multi-modal feature enhancement method is designed. By constructing a knowledge-guided attention mechanism, medical text knowledge is converted into prior weights and fused with data-driven attention weights. This guides the model to focus on key regions and features of diagnostic significance, thereby improving the medical consistency and interpretability of feature representations.Second, to address the challenges of fusion caused by multi-modal feature heterogeneity, we propose a multi-stream residual parallel fusion method. Through a multi-path residual structure, this method enables deep interaction and information compensation across modalities, effectively mitigating information loss and gradient propagation issues in traditional fusion methods, thereby enhancing overall feature expressiveness.Under a rigorous five-fold cross-validation on the renal tumor ultrasound dataset, the proposed KG-MMIL model achieves an accuracy of 91.6% and an AUC of 0.959. Compared with the second-best multi-modal method under the same evaluation protocol, the accuracy is improved by 1.4 percentage points. More importantly, the model exhibits the best stability across all evaluation metrics, with the smallest standard deviations, demonstrating strong robustness. These results validate that the proposed method successfully transitions from a 'black-box' to a clinically-aligned reasoning process by resolving modality heterogeneity and aligning with medical priors, representing a significant step towards interpretable medical AI.

Keywords: Multiple Instance Learning, multi-modal Fusion, Renal Tumor.

1 Introduction

Kidney disease is a major global public health issue, and its early and accurate diagnosis is crucial for improving patient prognosis [1]. Among various imaging modalities, 2D gray-scale [2] ultrasound and its Doppler techniques have become essential methods for the screening and diagnosis of kidney disease due to their non-invasive, real-time, and radiation-free advantages.Gray-scale ultrasound clearly displays the morphological structures of the kidneys (such as size and cortical thickness), while color Doppler and spectral Doppler provide valuable hemodynamic information, collectively serving as a crucial basis for clinical decision-making [3].

However, the accuracy of ultrasound diagnosis remains highly dependent on the individual experience and subjective judgment of the physician, posing significant challenges [4].First, differentiating benign from malignant solid renal masses (such as renal angiomyolipomas and renal cell carcinoma) proves particularly challenging, as their ultrasonographic features frequently exhibit considerable overlap (e.g., atypical renal cell carcinoma and angiomyolipomas may both appear hyperechoic), leading to a high risk of misdiagnosis or missed diagnosis.Second, manual measurements (such as tumor size and blood flow velocity) introduce human subjective errors, making it difficult to ensure diagnostic consistency [5].

Currently, deep learning-based methods for ultrasound-assisted diagnosis of renal tumors suffer from significant "black-box" [6] issues and performance bottlenecks. Most studies rely solely on single-modality images [7] and fail to effectively integrate hemodynamic information provided by Doppler techniques, resulting in an incomplete feature set and limited diagnostic performance.Even when attempting multi-modal fusion, existing strategies—such as simple feature concatenation or decision weighting—remain crude and constitute "black-box" operations. Furthermore, they fail to embed critical medical prior knowledge (such as the rule explicitly stated in clinical diagnostic guidelines that "hamartomas typically present as hyperechoic with reduced blood flow") into the model's inference pipeline. Consequently, the model's reasoning logic is severely disconnected from the diagnostic reasoning of clinicians, resulting in insufficient reliability and interpretability of the results.

To address the "black-box" issue, researchers in the field of explainable AI have proposed various methods.In terms of visual explanations, the Grad-CAM [8] technique proposed by Selvaraju et al. generates visual heatmaps using gradient information, enabling post-hoc visualization of the key image regions on which the model's decisions rely; regarding model architecture design, built-in interpretability methods such as attention mechanisms (e.g., as applied by Luo et al. in breast cancer analysis [9]) allow the model to prioritize critical features throughout the decision-making workflow.However, most of these methods fall under the category of "post-hoc explanations" for completed decisions, or remain purely data-driven, lacking medical rationality constraints [10].Therefore, our approach goes beyond simple application: we innovatively incorporate the core ideas of multi-instance learning frameworks and attention mechanisms, but our key breakthrough lies in proposing the conversion of medical text knowledge into computable prior signals. This aims to deeply integrate with data-driven attention,

thereby guiding the model to actively follow clinical logic during the decision-making process, achieving a leap from "explaining decisions" to "guiding decisions."

Combining the aforementioned techniques and inspired by the Multi-Instance Learning (MIL) [11] framework, this paper proposes a knowledge-guided multi-modal multi-instance learning model designed tailored for accurate and interpretable auxiliary diagnosis of renal tumors.The core idea of this model is to transform medical prior knowledge into learnable guidance signals that adaptively fuse multi-modal ultrasound features. By integrating medical knowledge on top of multi-modal fusion, the model enhances detection accuracy.Experiments demonstrate that KG-MMIL attains an accuracy of 91.6%, with an AUC of 0.960 on a renal tumor ultrasound dataset, outperforming baseline methods. Additionally, it generates heatmaps to provide intuitive explanations, effectively balancing effectiveness and transparency.

The key contributions of this study are summarized as follows:

Proposal of a knowledge-guided attention and fusion framework: We designed a novel attention mechanism capable of encoding medical text knowledge into prior weights and dynamically fusing them with data-driven attention. This mechanism renders the model's inference pipeline transparent while ensuring alignment with clinical diagnostic logic.

We constructed an efficient multi-modal feature integration and fusion pipeline: To address the heterogeneity of grayscale, color Doppler, and spectral Doppler data, we designed dedicated feature extractors and a residual parallel fusion module, effectively capturing complementary morphological and hemodynamic information to significantly improve diagnostic performance.

We have implemented interpretable heatmap-assisted diagnosis: the model not only outputs classification results but also generates visual heatmaps that intuitively display the key image regions underlying the decision, greatly enhancing the credibility of the results and their clinical utility.

2 Related Work

Early research primarily focused on classifying renal tumors using single grayscale ultrasound images, with typical methods relying on feature learning architectures leveraging convolutional neural networks (CNNs). The ResNet network proposed by He et al. [12] effectively mitigated the vanishing gradient problem in deep network training through residual connections and has been widely applied to medical image classification tasks.The DenseNet proposed by Huang et al. [13] enhances feature reuse capabilities through a dense connection mechanism, thereby improving the model's expressive power. Additionally, the EfficientNet proposed by Tan et al. [14] achieves a good balance between model accuracy and computational efficiency through a composite scaling strategy.In recent years, the Transformer architecture has also been introduced to visual tasks. The Vision Transformer (ViT), proposed by Dosovitskiy et al. [15], was the first to apply a pure Transformer architecture to image classification tasks, while

the Swin Transformer, proposed by Liu et al. [16], significantly improved model performance in visual tasks through a hierarchical window attention mechanism. These methods have achieved good results in single-modality medical image analysis and serve as important comparison baselines in the experiments of this paper.

To address the constraints of single-modality data, researchers have gradually introduced multi-modal learning into the domain of medical imaging, aiming to enhance the model's overall discriminative capability by fusing information from different sources (e.g., grayscale structural information and Doppler blood flow information). Based on the stage of fusion, existing multi-modal methods are typically categorized into three types: early fusion, mid-fusion, and late fusion [17]. Among these, early fusion approaches commonly employ concatenation (Concat) of multi-modal data at the input level or prior to feature extraction, then feed the combined data into the model for training.This method is simple to implement and can preserve some of the original cross-modal information to a certain extent. However, due to significant differences in distribution and semantic levels across modalities, direct fusion is prone to introducing noise interference, which limits model performance;Late fusion methods, on the other hand, train independent models for each modality and integrate the results at the decision stage through weighted averaging or voting mechanisms. These methods offer better stability and flexibility; however, due to the lack of interaction at the feature level, they struggle to fully exploit the complementary information between modalities;In contrast, intermediate fusion methods introduce cross-modal interaction mechanisms during feature extraction. By performing fusion at an intermediate feature layer—known as Intermediate Fusion—they can more effectively model the intermodal relationships and typically achieve superior performance. Building on this, some studies have further proposed specialized multi-modal network architectures, such as MCNet [18], which design cross-modal feature interaction paths to enhance information fusion capabilities and demonstrate good performance in medical imaging tasks.Although the aforementioned methods have made some progress in multi-modal medical image analysis, significant shortcomings remain: on the one hand, most fusion strategies rely on simple concatenation or weighted operations, making it difficult to capture complex nonlinear cross-modal relationships; on the other hand, these methods largely follow a purely data-driven paradigm, failing to explicitly incorporate medical prior knowledge into the models, which results in deficiencies in terms of interpretability and clinical consistency.Therefore, how to introduce medical knowledge constraints into the multi-modal fusion process and design more effective cross-modal feature interaction mechanisms has emerged as a pressing challenge requiring immediate resolution.

As the complexity of deep learning models increases, their "black-box" nature severely hinders their clinical implementation. Consequently, Explainable AI (XAI) technology has emerged. The Grad-CAM 8 technique proposed by Selvaraju et al. is the current industry standard; it generates heatmaps by backpropagating gradients, thereby visualizing the regions of interest in the model's inference pipeline.In the field of ultrasound, Eren et al 19. innovatively introduced attention mechanisms into breast ultrasound diagnostic networks. By leveraging spatial and channel-wise attention mechanisms, they directed the network to automatically concentrate on lesion boundaries,

significantly improving segmentation and classification performance. However, research on explainability for multi-modal renal ultrasound remains in its infancy.Most existing work considers explainability just as a 'post-hoc validation' tool after training, rather than actively utilizing clinical insights to modulate the model's attention amidst the computational process. Attention mechanisms lacking medical knowledge constraints may sometimes misinterpret background noise as key features, leading to unreliable diagnostic results. Therefore, constructing a framework that deeply integrates medical prior knowledge with data-driven attention is an urgent issue that needs to be addressed.

3 Methodology

3.1 Overview

To improve the overall diagnostic performance and interpretability of renal ultrasound analysis, we propose a Knowledge-Guided Multi-Modal Multi-Instance Learning (KG-MMIL) framework. This framework aims to fully leverage the morphological information from grayscale ultrasound, the blood flow distribution information from color Doppler, and the hemodynamic quantitative parameters from spectral Doppler.The overall framework follows an architectural design of "multi-view feature extraction—multi-level attention aggregation—residual parallel fusion—classification prediction," as shown in 错误!未找到引用源。.

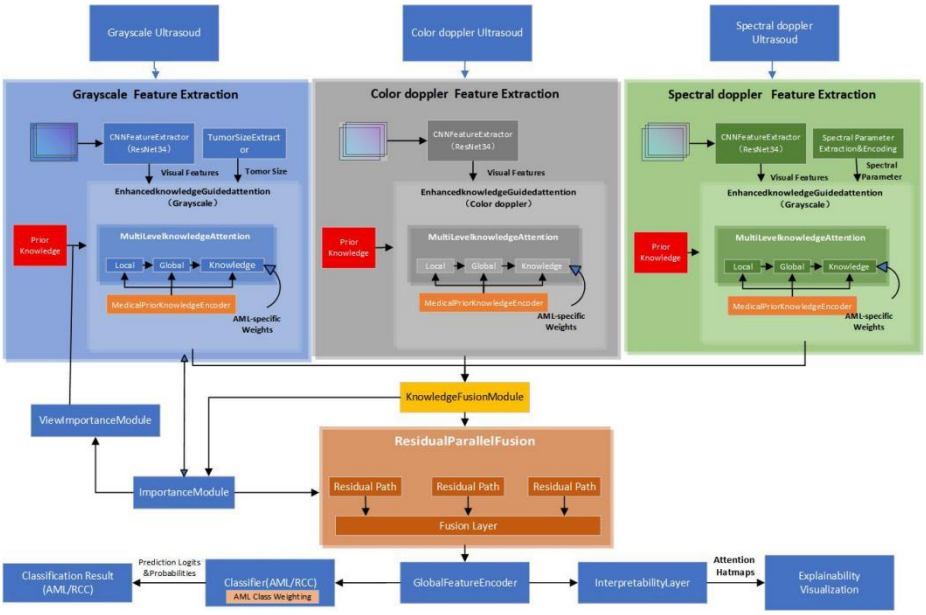


Fig. 1. Overall framework of KG-MMIL for multi-modal ultrasound diagnosis of renal tumors

Given multi-modal inputs (including grayscale, color Doppler, and spectral Doppler images), the data first passes through a feature extraction network based on ResNet34 to generate basic visual feature vectors.To simulate clinical quantitative diagnosis, we concurrently introduce a lightweight tumor size extractor and a spectral parameter encoder, which utilize image processing techniques and an MLP, respectively, to encode key indicators such as lesion diameter, systolic peak flow velocity, and resistance index into high-dimensional features.Next, these instance-level features are fed into a multi-level knowledge-guided attention module, which comprises three parallel branches—local, global, and prior knowledge—and, in conjunction with a view importance module, adaptively aggregates multi-view features into highly discriminative package-level feature representations.

In the feature fusion stage, unlike traditional simple concatenation, we introduce a residual parallel fusion module comprising multiple parallel enhancement paths. This module first performs nonlinear mapping on each modality's features via an independent Linear-ReLU-Dropout structure to preserve modality-specificity, followed by deep fusion through residual connections and average pooling.Finally, the fused features pass through a classification head incorporating an AML category reweighting strategy, outputting benign/malignant prediction probabilities and generating interpretable heatmaps for clinical validation.

3.2 Multi-level Knowledge-Guided Attention Mechanism

To effectively aggregate instance-level features while balancing local texture and global context, we propose a multi-level knowledge-guided attention mechanism, as illustrated in **错误!未找到引用源。**. It is designed to adaptively integrate features from three complementary dimensions, thereby accommodating the heterogeneity across different ultrasound modalities.

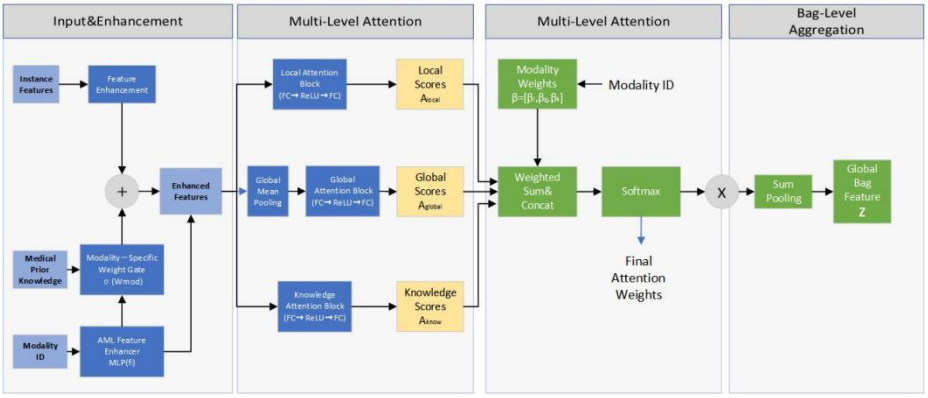


Fig. 2. Structure of the Multi-level Knowledge-Guided Attention Mechanism

Input instance features are denoted as $F_{in} \in R^{N \times D}$, comprising N instances with D-dimensional features. The data is processed in parallel through three independent branches to compute raw attention scores. The local attention branch utilizes a multi-layer perceptron MLP to capture fine-grained texture features; the global attention

branch first computes the global mean of the features $g = \mathrm{Mean}(F_{in})$, and then maps global contextual constraints via an MLP; the knowledge attention branch focuses on extracting feature patterns aligned with medical prior knowledge.This branch is implemented through a lightweight multilayer perceptron, structured as: Linear(D → D/2) → ReLU → Dropout(p=0.1) → Linear(D/2 → 1). The input instance features are initially projected into a lower-dimensional space to capture high-level semantic interactions, and are subsequently mapped to a scalar attention score. This design ensures nonlinear expressive capability while preventing overfitting to specific prior patterns via the Dropout layer, enabling the model to learn more robust feature representations pertinent to medical knowledge.This parallel computation process can be represented as:

$$A_{loc} = \phi_{loc}(F_{in}) \tag{1}$$

$$A_{glob} = \phi_{glob}(g \cdot 1^{T}) \tag{2}$$

$$A_{know} = \phi_{know}(F_{in}) \tag{3}$$

Where $\phi(\cdot)$ denotes a nonlinear mapping composed of linear layers and a ReLU activation function.

More specifically, to address the differing dependencies on feature dimensions across different modalities such as grayscale and color Doppler, we introduce learnable modality-specific weight vectors $w_m \in R^3$. For the current modality m, these weights are first activated via a Sigmoid function and normalized to generate adaptive fusion coefficients W_m. This mechanism allows the model to dynamically adjust the contribution ratios of the local, global, and knowledge branches, calculated as follows:

$$W_m = \frac{Sigmoid(w_m)}{\sum Sigmoid(w_m)} = [\alpha_{loc}, \alpha_{glob}, \alpha_{know}] \tag{4}$$

We define these normalized weights as the final attention weight for each instance, denoted as $A^{(i)}_{final}$, that is:

$$A^{(i)}_{final} = \mathrm{Softmax}\left(\alpha_{loc}A^{(i)}_{loc} + \alpha_{glob}A^{(i)}_{glob} + \alpha_{know}A^{(i)}_{know}\right) \tag{5}$$

Finally, the normalized fusion weights are element-wise weighted and summed with the attention maps from the three branches, then normalized at the instance level via the Softmax function to generate the final instance importance distribution. This distribution is used to perform weighted aggregation on the raw features, thereby generating a highly discriminative batch-level feature representation Z, completing the feature mapping from instances to batches:

$$Z = \sum_{i=1}^{N} A^{(i)}_{final} \cdot f_i \tag{6}$$

3.3 Multi-Stream Residual Parallel Fusion Module

To address the issue of inconsistent distribution of heterogeneous modalities (such as grayscale morphological features and spectral Doppler blood flow parameters) in the feature space, and to avoid dimensional redundancy caused by direct concatenation, we

designed a Multi-Stream Residual Parallel Fusion (RPF) module. This module adopts an "independent manifold mapping–residual aggregation" architecture, as shown in 错误!未找到引用源。:

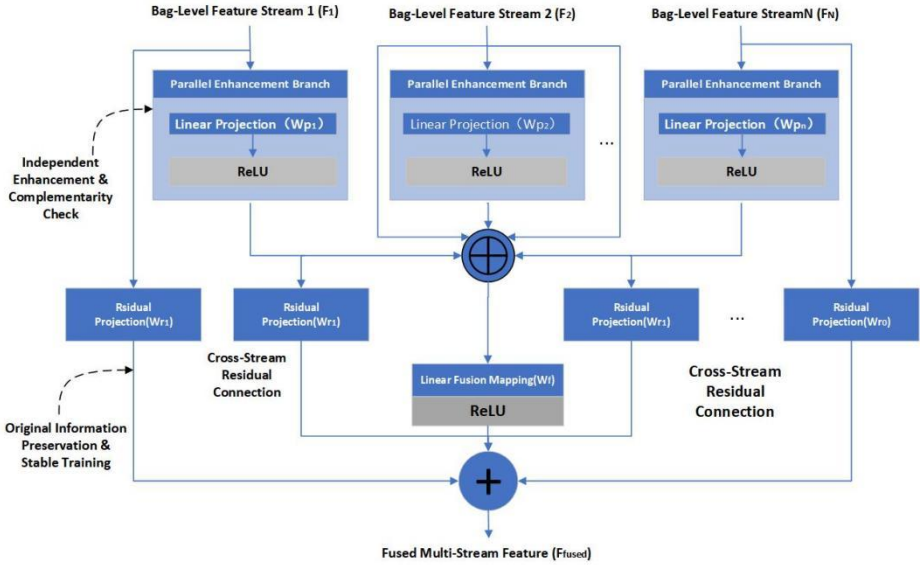


Fig. 3. Structure of the Multi-Stream Residual Parallel Fusion Module.

Specifically, this module receives three patch-level feature vectors from the upstream attention mechanism:Z_{gray} ,Z_{color} , andZ_{spec} . These three feature vectors are not immediately merged but instead enter three independent modal enhancement paths. Each path is essentially a deep residual network designed to project modality-specific features into a shared, highly discriminative semantic space. This parallel mapping process is defined as:

$$H_m = \varphi_m(Z_m),\ \ m \in \{gray, color, spec\} \tag{7}$$

Where $\varphi_m(\cdot)$ denotes a sequence of nonlinear transformations consisting of "Linear-ReLU-Dropout-Linear,"designed to extract high-order semantic features of the modality.

Furthermore, to promote complementarity among modalities and maintain training stability, we introduce a residual parallel aggregation strategy. Unlike traditional weighted concatenation, we first stack the enhanced features H_m and compute their centroid (i.e., the average feature) in the feature space.Simultaneously, to prevent gradient vanishing in deep networks and preserve the distribution assumptions of the original features, we add the mean of the original input features as a residual term to the fused features. This fusion operation can be expressed as:

$$F_{agg} = \frac{1}{3}\sum_m H_m + \lambda \cdot \left(\frac{1}{3}\sum_m Z_m\right) \tag{8}$$

where λ is a residual weight coefficient that regulates the contribution of the original features in the fusion result. In this paper, we optimize this hyperparameter via a grid search on the validation set, finally setting λ = 0.5. This value aims to achieve the optimal balance between preserving the fundamental distribution of the original features and leveraging the enhanced cross-modal interaction capabilities of the augmented features.

Mechanistically, a larger value of λ causes the fused output to favor retaining the original features, which helps maintain training stability. Conversely, a smaller λ relies more on the nonlinearly enhanced features, facilitating deeper cross-modal interaction. We further tested the performance of λ within the range of [0.3, 0.7] and observed that the fluctuation in model accuracy was less than 0.5%. This indicates that the proposed fusion strategy exhibits good robustness to variations in this parameter within a certain range.

Finally, the aggregated multi-modal featuresF_{agg} undergo dimension reduction and feature reorganization via a final linear projection layer, generating the final decision feature vectorV_{final} used for classification. This design not only achieves deep interaction of multi-modal information but also significantly enhances the model's robustness when handling missing or noisy data through the residual structure:

$$V_{final} = LayerNorm\left(Linear\left(F_{agg}\right)\right) \tag{9}$$

3.4 Dynamic Weighted Prediction and Interpretability

To improve the sensitivity of renal angiomyolipoma detection and enhance the model's clinical interpretability, we designed a dynamic weighting classification and visualization module.

Specifically, the fused feature vector V_{final} is fed into a fully connected classification head. After applying Dropout to prevent overfitting, a linear layer maps the input to generate the raw class prediction scores $L = [l_{aml},\ l_{rcc}]$.

To address training imbalance caused by the small number of AML samples, we design a learnable dynamic weighting mechanism. This mechanism introduces a trainable parameter τ, initialized to 0. This initialization is based on the property of the Sigmoid function: σ(0) = 0.5, which ensures that, at the start of training, the model's focus on the AML class is in a neutral state without imposing a prior bias, thereby guaranteeing adaptive optimization. During each forward pass, this parameter is transformed via the Sigmoid function into a dynamic weight for the AML class, which is used to scale the class's raw score:

$$l'_{aml} = l_{aml} \cdot \sigma(\tau) \tag{10}$$

$$P = \mathrm{Softmax}\left(\left[l'_{aml},\ l_{rcc}\right]\right) \tag{11}$$

The parameter τ is optimized via backpropagation and gradient descent, with the same objective as the overall model, namely minimizing the cross-entropy loss. This design enables adaptive weight adjustment: when the model encounters difficulty in classifying AML samples, the corresponding loss increases, and the resulting gradients drive τ

to increase accordingly. Consequently, this mechanism effectively assigns higher weights to AML samples in the loss function, encouraging the model to place greater emphasis on these samples during subsequent training.

Finally, to provide intuitive diagnostic evidence, we utilize the attention weights computed in Section 3.2 to generate class activation maps (CAM). By projecting high-weight feature regions back onto the original images, the model can highlight key lesion areas that contribute to the decision-making process, thereby offering visual interpretability and assisting clinicians with supportive evidence for diagnosis:

$$\mathrm{M}(\mathrm{x},\mathrm{y}) = \sum_{\mathrm{i}} \mathrm{A}_{\mathrm{final}}^{(\mathrm{i})} \cdot \mathbb{I}\big((\mathrm{x},\mathrm{y}) \in \mathrm{Patch}_{\mathrm{i}}\big) \tag{12}$$

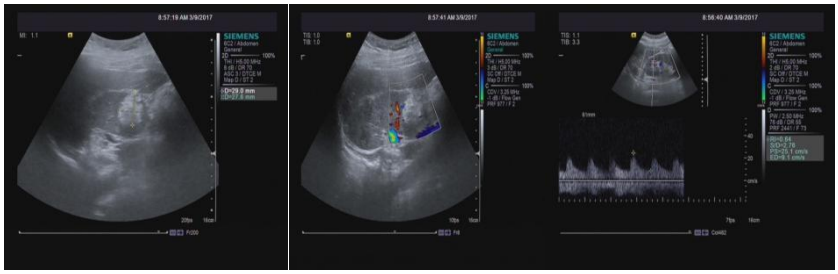
Here, $\mathrm{M}(\mathrm{x},\mathrm{y})$ is the value of the generated heatmap at coordinate (x,y), and $\mathbb{I}(\cdot)$ is an indicator function used to assign the attention weight of the i-th instance to its corresponding image patch.

4 Experiments

4.1 Dataset

This study is based on a renal tumor ultrasound imaging dataset derived from real clinical cases in a hospital. The dataset consists of 215 patient cases, categorized into 129 renal cell carcinomas (RCC) and 86 angiomyolipomas (AML). Each case encompasses three imaging modalities: grayscale ultrasound, color Doppler, and spectral Doppler. All images are standardized to a resolution of 512 × 512 pixels and saved in JPG format. To assess the stability and generalization capability of the proposed model, a five-fold cross-validation strategy was implemented. Specifically, the entire dataset was randomly partitioned into five mutually exclusive subsets. During each iteration, four subsets were utilized for training while the remaining one served as the test set. This procedure was iterated five times to ensure each subset was validated exactly once. The final results are reported as the mean performance accompanied by the standard deviation across all folds.

Furthermore, to enable a fair comparison with state-of-the-art methods, a fixed data split ratio of 7:1:2 (training:validation:testing) was also adopted for supplementary analysis. Representative examples from the dataset are illustrated in **错误!未找到引用源。**.



(a) (b) (c)

Fig. 4. Representative multi-modal ultrasound images of renal tumors, including (a) grayscale ultrasound, (b) color Doppler, (c) spectral Doppler.

4.2 Evaluation Metrics

To evaluate the performance of the proposed model, we adopt accuracy, F1-score, precision, recall, and the area under the receiver operating characteristic curve (AUC) as evaluation criteria. Specifically, accuracy quantifies the overall ratio of correctly classified instances; F1-score offers a harmonic mean that balances precision and recall; recall gauges the model's sensitivity in detecting positive samples; precision reflects the reliability of the model's positive predictions; and AUC serves as an indicator of the model's discriminative power between positive and negative classes.

4.3 Implementation Details

All experiments were conducted utilizing the PyTorch framework on a workstation equipped with an NVIDIA RTX 3090 GPU. During data preprocessing, all images were uniformly resized to 512 × 512 pixels and normalized to obtain zero mean and unit variance. To bolster the model's generalization capability, various data augmentation strategies were employed, including random horizontal flipping and random rotation (±15°).

The proposed multi-modal framework comprises three parallel feature extraction branches, where each modality is encoded via a lightweight convolutional backbone network. Subsequently, the extracted features are fed into a knowledge-guided attention module, which integrates medical prior knowledge to refine the model's focus on discriminative regions.

Under the multiple instance learning setting, each bag consists of a collection of instance-level features, and an attention-based aggregation strategy is adopted to derive the bag-level representation. The final classification prediction is generated through a fully connected layer coupled with a Softmax function.

The model was optimized using the cross-entropy loss function. During training, the Adam optimizer was utilized for parameter updates, with an initial learning rate of 1×10^{-4}, a weight decay coefficient of 1×10^{-5}, and β_1 and β_2 set to 0.9 and 0.999, respectively. The batch size was configured to 16, and the model was trained for 100 epochs. Furthermore, the learning rate was decayed to 0.1 times its previous value every 30 epochs.

In terms of computational efficiency, the proposed KG-MMIL model has approximately 21.3M parameters, requires about 16.8 GFLOPs, and processes a single image in approximately 23 ms on average. These results demonstrate that the model achieves high diagnostic accuracy while maintaining favorable computational efficiency, indicating promising potential for practical deployment.

Furthermore, to ensure fairness and comparability of the experimental results, all baseline methods are implemented and evaluated under the same training settings and hardware environment.

4.4 Comparative Experiments

To validate the effectiveness of the proposed method, we conduct a comprehensive comparison between the KG-MMIL model and a series of state-of-the-art unimodal and multi-modal baseline methods on the renal tumor ultrasound dataset. To enhance the rigor and statistical robustness of the evaluation, five-fold cross-validation is adopted as the primary evaluation strategy, with the results reported in **错误!未找到引用源。**. Meanwhile, to facilitate direct comparison with existing studies that use fixed data splits, additional experiments are conducted under a consistent 7:1:2 split, with the results presented in **错误!未找到引用源。**.

In the five-fold cross-validation setting, the entire dataset is randomly partitioned into five folds. Each fold is used once as the test set, while the remaining four folds are used for training (with 20% of the training data further split as a validation set). This process is repeated five times. **错误!未找到引用源。** reports the mean and standard deviation of the performance across all five folds, reflecting both the average performance and stability of the model.

The comparison methods include: (1) unimodal approaches, consisting of classical CNN-based models using grayscale ultrasound images (ResNet-50, DenseNet-121, EfficientNet-B4) and vision Transformer-based models (ViT-Base, Swin Transformer); and (2) multi-modal fusion approaches, including early fusion (Concat), late fusion (Late Fusion), intermediate fusion (Intermediate Fusion), and a dedicated multi-modal network, MCNet.

Table 1. Classification Metrics Under Different Cross-Validation Strategies (Mean ± Std)

Experiment	Method	Accuracy	Precision	Recall	F1-score	AUC
Single-modal	ResNet-50	0.851 ± 0.012	0.844 ± 0.015	0.852 ± 0.018	0.848 ± 0.012	0.92 ± 0.006
	DenseNet-121	0.859 ± 0.011	0.852 ± 0.013	0.860 ± 0.016	0.856 ± 0.011	0.928 ± 0.006
	EfficientNet-B4	0.865 ± 0.010	0.858 ± 0.012	0.864 ± 0.015	0.861 ± 0.011	0.932 ± 0.005
	ViT-Base	0.852 ± 0.013	0.848 ± 0.016	0.850 ± 0.017	0.849 ± 0.013	0.923 ± 0.007
	Swin-Transformer	0.863 ± 0.010	0.856 ± 0.012	0.862 ± 0.015	0.859 ± 0.010	0.930 ± 0.005
Multi-modal	Late Fusion	0.898 ± 0.011	0.878 ± 0.013	0.872 ± 0.014	0.875 ± 0.011	0.952 ± 0.005
	Intermediate Fusion	0.905 ± 0.010	0.885 ± 0.011	0.877 ± 0.013	0.881 ± 0.010	0.954 ± 0.004
	MCNet	0.880 ± 0.012	0.860 ± 0.014	0.855 ± 0.016	0.857 ± 0.013	0.937 ± 0.006
	Concat	0.902 ± 0.009	0.882 ± 0.010	0.874 ± 0.012	0.878 ± 0.009	0.955 ± 0.004
Ours	KG-MMIL	0.916 ± 0.008	0.895 ± 0.005	0.878 ± 0.005	0.884 ± 0.005	0.959 ± 0.003

Table 2. Quantitative evaluation of different models under the 7:1:2 data split

Experiment	Method	Accuracy	Precision	Recall	F1-score	AUC
Single-modal	ResNet-50	0.826	0.814	0.836	0.824	0.893
	DenseNet-121	0.838	0.829	0.841	0.834	0.902
	EfficientNet-B4	0.851	0.846	0.857	0.851	0.919
	ViT-Base	0.847	0.852	0.831	0.841	0.916
	Swin-Transformer	0.860	0.864	0.859	0.861	0.929

Multi-modal	Late Fusion	0.879	0.861	0.858	0.859	0.939

Table 2 .(continue).

Experiment	Method	Accuracy	Precision	Recall	F1-score	AUC
	Intermediate Fusion	0.893	0.878	0.872	0.875	0.949
Multi-modal	MCNet	0.853	0.812	0.798	0.805	0.912
	Concat	0.902	0.891	0.883	0.887	0.962
Ours	KG-MMIL	0.919	0.896	0.878	0.887	0.961

Under the five-fold cross-validation framework, the KG-MMIL model achieves the best overall diagnostic performance, with an average accuracy and AUC of 91.6% and 0.959, respectively. Compared with the second-best multi-modal method (Concat, accuracy 90.2%), the proposed model improves accuracy by 1.4 percentage points. More importantly, it exhibits the smallest standard deviation across all evaluation metrics, indicating that its performance is least sensitive to variations in data partitioning and demonstrating strong stability and generalization capability.

For comparison, experiments conducted under the same fixed split (7:1:2) show that KG-MMIL also achieves superior performance (accuracy 91.9%, AUC 0.961). This result is consistent with the findings from cross-validation, jointly validating the effectiveness of the proposed method from multiple evaluation perspectives.

4.5 Ablation Experiments

In order to empirically verify the necessity of individual essential parts in the KG-MMIL architecture under a statistically rigorous setting, we carried out systematic comparative experiments. By adopting a five-fold cross-validation protocol, we aimed to delineate the specific improvements attributed to each part, considering both the model's overall performance and its robustness. Statistical summaries, including the mean and standard deviation, for all metrics are presented.

We design four key model variants for stepwise comparative analysis: (A) baseline model, which removes both the knowledge-guided attention mechanism and the AML dynamic weighting module; (B) model A with the addition of the knowledge-guided attention module; (C) model A with the addition of the AML dynamic weighting module; and (D) the full KG-MMIL model, which incorporates both knowledge-guided attention and AML dynamic weighting. Detailed five-fold cross-validation results are presented in **错误!未找到引用源。**.

Table 3. Quantitative analysis of core components via five-fold cross-validation.

Model Variant	Accuracy	Precision	Recall	F1-score	AUC	AML Recall
A.Baseline	0.902 ± 0.009	0.882 ± 0.010	0.874 ± 0.012	0.878 ± 0.009	0.955 ± 0.004	0.865 ± 0.015
B.A+Multi-stream Residual Fusion	0.908 ± 0.008	0.888 ± 0.009	0.876 ± 0.011	0.882 ± 0.008	0.957 ± 0.004	0.870 ± 0.014
C.B+Knowledge-guided Attention	0.912 ± 0.007	0.892 ± 0.007	0.878 ± 0.009	0.885 ± 0.007	0.958 ± 0.003	0.872 ± 0.012
D. Full KG-MMIL	0.916 ± 0.008	0.895 ± 0.006	0.878 ± 0.008	0.886 ± 0.006	0.960 ± 0.003	0.875 ± 0.010

The ablation results in 错误!未找到引用源。 demonstrate that each core component effectively enhances the model's performance and stability, with mechanisms that directly address the challenges outlined in the Introduction. First, the Multi-stream Residual Parallel Fusion module increases accuracy from 90.2% to 90.8%. By preserving the specificity of grayscale morphological features and spectral parameters through independent mapping pathways and utilizing a residual connection (λ=0.5) to prevent gradient vanishing, it resolves the feature conflicts and information loss caused by modal heterogeneity. Building on this, the Knowledge-Guided Attention mechanism further elevates accuracy to 91.2%, mitigating the limitations of pure data-driven "black-box" models. Rather than simply stacking features, this mechanism encodes clinical priors into learnable prior vectors and adaptively fuses them with data-driven attention scores via Eq. (5), forcing the model to prioritize pathologically critical regions during the feature aggregation stage. This achieves the transition from "post-hoc explanation" to "process guidance." Finally, the AML Dynamic Weighting mechanism improves the AML recall to 0.875 without sacrificing overall precision. The underlying mechanism is that τacts as a learnable scaling factor; during backpropagation, it automatically increases the penalty weight for AML samples (Eq. 10-11). When the model struggles to identify such features, the loss function drives $\tau$to increase, forcing the classifier to maintain higher sensitivity for these features.

Finally, the full model integrating all components achieves the best overall performance (average accuracy of 91.6% and AUC of 0.960), while maintaining the best stability across most key metrics, demonstrating the synergistic effect among all components.

To explicitly evaluate the independent contributions of each ultrasound imaging modality in renal tumor diagnosis and to provide guidance for prioritizing clinical data acquisition, we design a systematic unimodal ablation study. All experiments in this section are conducted under the five-fold cross-validation setting to assess the performance of four modality input combinations: (1) grayscale ultrasound only (Gray-only); (2) grayscale combined with color Doppler (Gray+Color); (3) grayscale combined with spectral Doppler (Gray+Spectral); and (4) full three-modality input (Gray+Color+Spectral, i.e., KG-MMIL). The results are presented in 错误!未找到引用源。.

Table 4. Contribution analysis of individual modalities based on five-fold cross-validation

Modality Combination	Accuracy	Precision	Recall	F1-score	AUC
Gray-scale only	0.863 ± 0.010	0.856 ± 0.012	0.862± 0.015	0.859 ± 0.010	0.930 ± 0.005
Gray + Color Doppler	0.895 ± 0.009	0.882 ± 0.010	0.875 ± 0.012	0.878 ± 0.009	0.952 ± 0.004
Gray + Spectral Doppler	0.888 ± 0.010	0.876 ± 0.011	0.868 ± 0.013	0.872 ± 0.010	0.947 ± 0.005
Gray + Color + Spectral	0.916 ± 0.008	0.895 ± 0.005	0.878 ± 0.005	0.886 ± 0.005	0.960 ± 0.003

The unimodal ablation results in 错误!未找到引用源。 quantitatively characterize the diagnostic contributions of different modalities and elucidate the underlying physiological mechanisms. Using grayscale ultrasound alone, the model achieves an accuracy of 86.3%, confirming the fundamental value of morphological information. However, the introduction of Color Doppler yields the most significant performance

jump (89.5%, +3.2 percentage points). This gain directly addresses the "overlapping sonographic appearances" challenge outlined in the Introduction, as hemodynamic information provides the key evidence to differentiate hypervascular tumors from hypovascular ones. Furthermore, the addition of Spectral Doppler provides incremental gains (88.8%), with its quantitative parameters affording greater diagnostic objectivity. Finally, the three-modality fusion achieves the best performance and stability (91.6%, AUC 0.960, with the lowest standard deviations), demonstrating the strong complementarity and synergistic value of multi-modal information. Based on these findings, clinical data acquisition should prioritize complete three-modality imaging; when resources are constrained (e.g., in primary care), the combination of grayscale and Color Doppler represents an optimal trade-off between efficiency and diagnostic efficacy.

5 Limitations

Despite the promising performance of the proposed KG-MMIL model, several limitations still exist, mainly from four aspects: data, knowledge, model design, and clinical translation.

Limitations in Data and Knowledge Coverage. The dataset used in this study is collected from a single center and has a relatively limited scale, which restricts the diversity of training samples. In addition, the embedded medical prior knowledge mainly focuses on typical imaging characteristics, while lacking sufficient representation of different renal cell carcinoma subtypes and atypical angiomyolipoma. This limits the model's discriminative capability and generalization performance in complex or rare cases.

Limitations in Clinical Interpretability and Integration. Although the model provides visual heatmaps and knowledge-guided attention, representing a significant improvement over pure 'black-box' baselines, the interpretability remains primarily phenomenological rather than mechanistic. Currently, the model highlights 'where' (regions of interest) and aligns with 'what' (medical priors like hyperechoic), but it fails to explicitly link these features to underlying pathophysiological mechanisms. Beyond the algorithmic challenges, a major barrier to practical application is the seamless integration of model outputs into radiologists' existing diagnostic workflows. How to enable effective human–AI collaborative decision-making without increasing cognitive load remains a critical challenge for broader clinical adoption.

Future Directions.Future work will focus on the following directions: (1) collecting large-scale, multi-center datasets to validate generalization capability; (2) constructing a more fine-grained and structured ultrasound knowledge base for renal tumors to improve adaptability to rare subtypes; (3) exploring modality completion or robust fusion strategies to handle incomplete clinical inputs; and (4) conducting prospective clinical studies to evaluate the model's impact on diagnostic efficiency and accuracy in real-world settings.

6 Conclusion

This paper proposes a knowledge-guided multi-modal multiple instance learning model (KG-MMIL) for ultrasound-based renal tumor diagnosis. The model consists of two core components. First, a knowledge-guided attention mechanism encodes medical prior knowledge into learnable guidance signals, enabling the model to focus precisely on key imaging features that align with diagnostic reasoning, thereby enhancing the transparency and reliability of the decision-making process. Second, a multi-stream residual parallel fusion module is designed, which leverages independent modality-specific enhancement pathways and residual aggregation strategies to achieve deep and robust feature interactions among grayscale, color Doppler, and spectral Doppler modalities,improving the efficiency of multi-modal information utilization.

Under a more rigorous five-fold cross-validation framework, the KG-MMIL model achieves an average accuracy of 91.6% and an AUC of 0.960. Compared with the best unimodal baseline (Swin Transformer), the proposed method improves accuracy by 5.3 percentage points, while also maintaining a consistent performance advantage over the second-best multi-modal fusion approach. More importantly, the model exhibits the smallest standard deviations across all evaluation metrics, demonstrating excellent stability and generalization capability, and effectively addressing concerns regarding model robustness. In addition, the visualization heatmaps generated by the model provide intuitive evidence for its decision-making process, further satisfying clinical requirements for interpretability.

In summary, by deeply integrating medical knowledge into both the learning and inference processes and adopting a more rigorous evaluation protocol, this work not only improves the diagnostic accuracy of renal tumor ultrasound analysis but also significantly enhances the reliability and interpretability of the model, offering valuable insights and practical guidance for the development of clinically applicable intelligent diagnostic systems.

DATA AVAILABILITY

The KTMUS-D dataset used in this study was provided by the ZZCH. Due to data privacy policies, this dataset is not publicly available.

ACKNOWLEDGMENT

Project supported by the Natural Science Foundation of Hunan Province, China (Grant No. 2025JJ70030).

References

1. Jager, K.J., Kovesdy, C., Langham, R., et al.: A single number for advocacy and communication—worldwide more than 850 million individuals have kidney diseases. Nephrology Dialysis Transplantation 34(11), 1803–1805 (2019)
2. Gulati, M., Cheng, J., Loo, J.T., et al.: Pictorial review: Renal ultrasound. Clinical Imaging 51, 133–154 (2018)
3. Correas, J.M., Anglicheau, D., Joly, D., et al.: Ultrasound-based imaging methods of the kidney—recent developments. Kidney International 90(6), 1199–1210 (2016)
4. Liang, X., Du, M., Chen, Z.: Artificial intelligence-aided ultrasound in renal diseases: a systematic review. Quantitative Imaging in Medicine and Surgery 13(6), 3988–4002 (2023)
5. Wang, C., Yu, C., Yang, F., et al.: Diagnostic accuracy of contrast-enhanced ultrasound for renal cell carcinoma: a meta-analysis. Tumor Biology 35(7), 6343-6350 (2014)
6. Zhang, M., Ye, Z., Yuan, E., et al.: Imaging-based deep learning in kidney diseases: recent progress and future prospects. Insights into imaging 15(1), 50 (2024)
7. Schulz, S., Woerl, A.C., Jungmann, F., et al.: Multimodal deep learning for prognosis prediction in renal cancer. Frontiers in oncology 11, 788740 (2021)
8. Selvaraju, R.R., Cogswell, M., Das, A., et al.: Grad-CAM: visual explanations from deep networks via gradient-based localization. International journal of computer vision 128(2), 336-359 (2020)
9. Luo, Y., Huang, Q., Li, X.: Segmentation information with attention integration for classification of breast tumor in ultrasound image. Pattern Recognition 124, 108427 (2022)
10. Reyes, M., Meier, R., Pereira, S., et al.: On the interpretability of artificial intelligence in radiology: challenges and opportunities. Radiology: artificial intelligence 2(3), e190043 (2020)
11. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning, pp. 2127-2136. PMLR (2018)
12. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition, pp. 770-778 (2016)
13. Huang, G., Liu, Z., Van Der Maaten, L., et al.: Densely connected convolutional networks. In: IEEE conference on computer vision and pattern recognition, pp. 4700-4708 (2017)
14. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning, pp. 6105-6114. PMLR (2019)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
16. Liu, Z., Lin, Y., Cao, Y., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE/CVF international conference on computer vision, pp. 10012-10022 (2021)
17. Atrey, P.K., Hossain, M.A., El Saddik, A., et al.: Multimodal fusion for multimedia analysis: a survey. Multimedia systems 16(6), 345-379 (2010)
18. Guo, Z., Li, X., Huang, H., et al.: Medical image segmentation based on multi-modal convolutional neural network: Study on image fusion schemes. In: IEEE 15th International Symposium on Biomedical Imaging, pp. 903-907 (2018)
19. Eren, B.: Attention-guided residual learning with SEARNet for improved breast ultrasound lesion segmentation. Signal, Image and Video Processing 19(14), 1266 (2025)