International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 06 | Jun 2026 www.irjet.net p-ISSN: 2395-0072

CMR-FORMER: KNOWLEDGE-GUIDED MULTIMODAL REASONING TRANSFORMER FOR INTERPRETABLE ULTRASOUND LIVER LESION DIAGNOSIS

Srilega A[1], Mrs. H.S.Anuja[2]

[1]PG Scholar Bio-Medical Department Udaya School of engineering, Kanyakumari, Tamil Nadu, India.
[2]Assistant Professor Bio-Medical Department, Udaya School of engineering, Kanyakumari, Tamil Nadu, India.

Abstract-*Ultrasound imaging is widely used for liver lesion diagnosis because it is non-invasive and provides real-time imaging. However, diagnosis is often time-consuming and dependent on operator expertise, while existing deep learning methods mainly rely on image data alone and lack clinical reasoning. To address these limitations, this study proposes CaMR-Former, a knowledge-guided multimodal reasoning transformer that combines ultrasound image features with clinical knowledge priors. The framework incorporates lesion-aware attention mechanisms, dual-path learning for classification and diagnostic reasoning, and an uncertainty-aware prediction module to improve diagnostic reliability. Additionally, attention visualization enhances explain ability by highlighting clinically relevant regions. Experimental results demonstrate that CaMR-Former provides accurate, interpretable, and trustworthy liver lesion diagnosis, supporting improved clinical decision-making.*

Keywords: Ultrasound Liver Lesion Diagnosis, Knowledge-Guided Multimodal Learning, Reasoning Transformer, Uncertainty-Aware Prediction, Explainable Artificial Intelligence (XAI).

1. INTRODUCTION

Ultrasound imaging is one of the most widely used diagnostic modalities for liver lesion assessment due to its non-invasive nature, real-time imaging capability, low cost, and widespread availability. It plays a crucial role in the early detection and characterization of focal liver lesions (FLLs), including benign and malignant tumors. However, accurate interpretation of ultrasound images remains challenging because of speckle noise, low contrast, operator dependency, and substantial variability in lesion appearance. Recent studies have demonstrated that advanced deep learning techniques can improve diagnostic accuracy by automatically extracting discriminative features from ultrasound data, thereby supporting clinicians in liver lesion diagnosis [1], [2].

Most existing computer-aided diagnosis systems for liver lesion classification primarily rely on unimodal imaging information and often neglect complementary clinical knowledge that is routinely used by radiologists during decision-making. Clinical factors such as patient history, lesion morphology, enhancement patterns, and laboratory findings provide valuable contextual information for diagnosis. Several recent multimodal learning studies have shown that integrating ultrasound images with clinical information significantly enhances lesion classification performance and diagnostic reliability compared to image-only approaches [1], [4], [6].

To effectively capture complex lesion characteristics, transformer-based architectures have gained increasing attention in medical image analysis. Unlike conventional convolutional neural networks (CNNs), transformers employ self-attention mechanisms that can model long-range dependencies and global contextual relationships within medical images. Recent transformer-based frameworks, including SDR-Former and other multimodal diagnostic networks, have demonstrated superior performance in liver lesion classification, segmentation, and feature representation learning by leveraging both local and global image information [5], [8], [10].

Another important challenge in liver lesion diagnosis is the lack of interpretability in deep learning models. Clinical adoption of artificial intelligence systems requires transparent reasoning processes that enable physicians to understand and trust model predictions. Explainable artificial intelligence (XAI) techniques, such as attention visualization and feature attribution methods, have been increasingly incorporated into diagnostic frameworks to highlight clinically relevant regions and provide interpretable decision support. Recent studies have shown that explainable multimodal architectures can improve clinician confidence while maintaining high diagnostic accuracy [8], [9], [10].

Furthermore, uncertainty estimation has emerged as a critical component of safe and reliable medical AI systems.

Diagnostic ambiguity is common in ultrasound imaging due to image artifacts, overlapping lesion characteristics, and limited image quality. Conventional classification models often generate highly confident predictions even for uncertain cases, potentially leading to diagnostic errors. Recent research has focused on uncertainty-aware learning strategies and confidence estimation mechanisms that can identify ambiguous cases and support risk-aware clinical decision-making, thereby enhancing patient safety and model robustness [2], [6], [9].

Therefore, this study proposes CaMR-Former, a Knowledge-Guided Multimodal Reasoning Transformer for interpretable ultrasound liver lesion diagnosis. The proposed framework integrates ultrasound image features with clinical knowledge priors through a multimodal attention mechanism and incorporates a lesion-aware token attention module to enhance abnormal region discrimination. Additionally, a dual-path learning strategy is employed for simultaneous lesion classification and diagnostic reasoning, while an uncertainty-aware prediction head estimates confidence levels for reliable decision support. Attention visualization further improves explainability by aligning model reasoning with clinically meaningful lesion regions. By combining multimodal learning, transformer-based reasoning, interpretability, and uncertainty estimation, the proposed approach aims to provide accurate, trustworthy, and clinically applicable liver lesion diagnosis, ultimately supporting improved patient management and treatment planning [1]–[10].

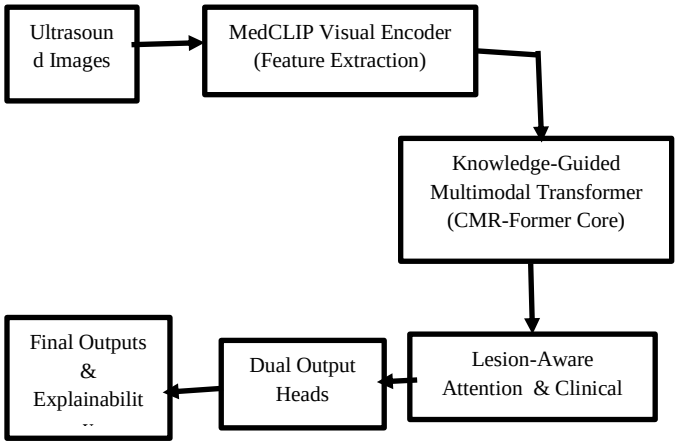

Fig.1 Architecture diagram

2. METHODOLOGIES

The proposed methodology is based on a Knowledge-Guided Multimodal Reasoning Transformer (CaMR-Former) for automated liver lesion diagnosis using ultrasound images. Initially, ultrasound images are provided as input and processed through a MedCLIP-based visual encoder to extract discriminative image features. These features capture important lesion characteristics such as texture patterns, lesion boundaries, and structural abnormalities, enabling effective representation of liver lesions for subsequent analysis.

To enhance diagnostic understanding, the extracted visual features are combined with clinical knowledge priors through a Knowledge Graph Attention Module. Medical attributes including lesion shape, echogenicity, and boundary information are incorporated to provide clinically meaningful context. Furthermore, a lesion-aware attention mechanism is employed to emphasize pathological regions while reducing the influence of irrelevant background tissues, thereby improving feature quality and diagnostic precision.

The integrated multimodal features are then processed using a dual-path learning framework. One branch performs liver lesion classification, while the second branch focuses on clinical reasoning and explanation generation. A ReportGPT-based decoder generates structured diagnostic reports that align with radiological reporting standards. Finally, an uncertainty-aware prediction head estimates confidence scores for each diagnosis, enabling reliable assessment of prediction certainty. This comprehensive framework provides accurate, interpretable, and clinically trustworthy liver lesion diagnosis by combining multimodal learning, reasoning, explainability, and uncertainty estimation.

2.1 Ultrasound Image Input & Preprocessing Module

The Ultrasound Image Input and Preprocessing Module serves as the initial stage of the proposed CaMR-Former framework. It receives both B-mode ultrasound and contrast-enhanced ultrasound (CEUS) images and prepares them for deep learning analysis. To reduce variations caused by different imaging devices and acquisition settings, intensity normalization is applied, ensuring consistent pixel distributions and improving the reliability of subsequent feature extraction.

To enhance image quality, anisotropic diffusion filtering is employed to suppress speckle noise while preserving important anatomical boundaries and lesion structures. This filtering technique improves the visibility of liver lesions and maintains critical edge information required for accurate diagnosis. Additionally, adaptive contrast enhancement is performed to highlight subtle differences

between healthy and abnormal tissues, making lesion regions more distinguishable.

After enhancement, all images are resized and standardized to a uniform resolution for efficient processing within the transformer architecture. The preprocessed images are then converted into patch-based representations, where each image is divided into fixed-size patches. This representation enables the transformer model to capture both local lesion characteristics and global spatial relationships, providing a structured and informative input for multimodal feature learning and diagnostic reasoning.

2.2 MedCLIP-Based Visual Encoder

The MedCLIP-Based Visual Encoder is responsible for extracting meaningful and clinically relevant features from liver ultrasound images. It employs a MedCLIP-based contrastive learning framework that aligns visual features with medical text descriptions, enabling the model to learn semantic relationships between imaging patterns and clinical findings. This approach enhances the interpretability and diagnostic relevance of the extracted features.

To effectively analyze liver lesions, the encoder utilizes a Vision Transformer (ViT) architecture that divides images into patches and processes them using self-attention mechanisms. Unlike conventional convolutional networks, the transformer captures both local lesion characteristics and global contextual information, allowing it to identify subtle abnormalities, tissue variations, and spatial relationships that are important for accurate diagnosis.

The encoder is further optimized through domain-specific medical pretraining, enabling robust performance in challenging ultrasound environments characterized by speckle noise, low contrast, and heterogeneous textures. The final output consists of discriminative visual tokens containing high-level semantic information about liver structures and abnormalities. These visual embeddings serve as the foundation for subsequent modules, including multimodal reasoning, lesion classification, and diagnostic report generation within the CaMR-Former framework.

2.3 Clinical Knowledge Graph Attention Module

The Clinical Knowledge Graph Attention Module integrates structured medical knowledge into the CaMR-Former framework to enhance diagnostic reasoning. A clinical knowledge graph is constructed using standardized medical guidelines and expert-defined relationships, representing important liver lesion characteristics such as morphology, echogenicity, margin definition, and internal texture. This structured representation enables the model to capture meaningful associations between clinical attributes and disease patterns.

To effectively learn from the knowledge graph, Graph Attention Networks (GAT) are employed to assign adaptive attention weights to different nodes and relationships. This attention mechanism allows the model to emphasize clinically significant features while reducing the influence of less relevant information. As a result, the generated knowledge embeddings provide rich contextual information that closely reflects expert diagnostic reasoning.

The extracted clinical knowledge embeddings are then integrated with visual features obtained from ultrasound images, enabling a multimodal understanding of liver lesions. By combining data-driven image analysis with expert-defined clinical priors, the framework supports hybrid symbolic–neural reasoning, improving diagnostic accuracy, consistency, and interpretability. This integration ensures that predictions are guided by both imaging evidence and established medical knowledge, leading to more reliable clinical decision support.

2.4 Lesion-Aware Token Attention Module

The Lesion-Aware Token Attention Module is designed to enhance the ability of the CaMR-Former framework to focus on clinically significant lesion regions within ultrasound images. It identifies lesion-specific visual tokens by computing attention scores and assigning higher importance to regions that are likely to contain abnormal tissue characteristics. This selective attention mechanism enables the model to prioritize lesion-related information

while reducing the influence of irrelevant image content.
To improve feature quality, the module employs self-attention mechanisms that suppress background interference caused by speckle noise, heterogeneous textures, and surrounding healthy tissues. By dynamically adjusting attention weights, the system emphasizes pathological regions and captures meaningful spatial relationships between lesions and adjacent liver structures. This contextual understanding supports more accurate characterization of lesion boundaries and internal tissue patterns.

Furthermore, the module enables implicit lesion localization without requiring explicit segmentation masks, reducing dependence on labor-intensive pixel-level

annotations. Through attention-driven learning, it automatically highlights suspicious regions and generates highly discriminative feature representations. As a result, the framework can effectively distinguish between benign and malignant lesions, improving diagnostic accuracy, robustness, and reliability in ultrasound-based liver lesion analysis.

2.5 Multimodal Reasoning Transformer (CMR-Former Core)

The Multimodal Reasoning Transformer (CMR-Former Core) acts as the central reasoning engine of the proposed framework, integrating visual information from ultrasound images with structured clinical knowledge. It combines lesion-aware visual tokens and clinical knowledge graph embeddings into a unified multimodal representation, enabling the model to analyze imaging findings alongside expert-defined medical concepts. This fusion enhances diagnostic understanding and provides a more comprehensive assessment of liver lesions.

To facilitate effective interaction between modalities, the transformer employs cross-attention mechanisms that allow visual features and clinical knowledge embeddings to influence each other. Through this bidirectional information exchange, the model learns meaningful semantic relationships between ultrasound patterns and clinical concepts, such as associating irregular lesion boundaries or hypoechoic regions with specific pathological conditions. This multimodal reasoning capability enables the framework to move beyond simple image recognition and perform clinically informed decision-making.

The module further utilizes stacked transformer blocks to capture both local and global contextual dependencies across image and knowledge representations. These deep attention layers generate interpretable multimodal embeddings that contain rich diagnostic information for subsequent analysis. The resulting representations serve as the foundation for lesion classification, clinical reasoning, and report generation, ensuring that the CaMR-Former framework delivers accurate, context-aware, and clinically aligned liver lesion diagnosis.

2.6 Dual-Path Learning Module

The Dual-Path Learning Module enables the CaMR-Former framework to simultaneously perform liver lesion classification and clinical reasoning. It divides the learning process into two complementary branches while utilizing a shared feature representation. This design ensures that the system not only predicts lesion categories accurately but also develops an understanding of the clinical factors contributing to its decisions.

The first branch, Path-1, focuses on liver lesion classification by learning discriminative patterns from multimodal features derived from ultrasound images and clinical knowledge embeddings. The second branch, Path-2, is dedicated to clinical reasoning and semantic alignment, where relationships between visual features and medical concepts are learned. This reasoning pathway helps the model associate imaging characteristics, such as lesion shape, texture, and echogenicity, with clinically relevant diagnostic interpretations.

Both branches share a common feature space, allowing knowledge gained from one task to enhance the performance of the other. Through joint optimization, the framework achieves a balance between diagnostic accuracy and interpretability while reducing the risk of overfitting. As a result, the Dual-Path Learning Module improves the robustness, reliability, and clinical applicability of the CaMR-Former framework for automated liver lesion diagnosis.

2.7 ReportGPT Diagnostic Report Generation Module

The ReportGPT Diagnostic Report Generation Module converts the multimodal reasoning outputs of the CaMR-Former framework into structured and clinically meaningful diagnostic reports. It utilizes a transformer-based language model trained on medical terminology and radiology reporting patterns to generate coherent, context-aware descriptions of liver lesions. This enables the system to provide automated reports that closely resemble those produced by experienced radiologists.

The report generation process is conditioned on multimodal information obtained from ultrasound images, clinical knowledge graphs, and transformer-based reasoning outputs. By integrating these diverse sources of information, the module produces accurate and clinically relevant descriptions that reflect lesion characteristics, diagnostic findings, and medical interpretations. This multimodal approach ensures that the generated reports are both informative and aligned with the diagnostic evidence.

To enhance usability in clinical settings, the module generates structured and guideline-compliant reports containing sections such as observations, impression, and conclusion. It also maintains consistency in medical terminology, reducing ambiguity and improving communication between AI systems and healthcare

professionals. As a result, the ReportGPT module supports efficient clinical documentation while improving the interpretability and practical applicability of the CaMR-Former framework.

2.8 Uncertainties-Aware Prediction Head

The Uncertainty-Aware Prediction Head is a safety-focused component of the CaMR-Former framework that evaluates the reliability of diagnostic predictions. Rather than providing only classification results, the module estimates predictive uncertainty using probabilistic techniques such as Monte Carlo Dropout or Bayesian inference. By performing multiple inference passes, the system measures prediction variability and determines the confidence level associated with each diagnosis.

Based on the probabilistic outputs, the module generates confidence scores that indicate the certainty of liver lesion classification results. High confidence values reflect strong agreement among model predictions, while lower confidence values suggest uncertainty or ambiguity in the diagnostic outcome. These confidence estimates provide valuable information for assessing the trustworthiness of automated predictions in clinical practice.

The module further identifies and flags low-confidence cases for expert review, enabling a human-in-the-loop decision-making process. This prevents uncertain predictions from being directly used in critical clinical decisions and reduces the risk of misdiagnosis. By integrating uncertainty estimation into the diagnostic workflow, the Uncertainty-Aware Prediction Head enhances clinical safety, transparency, and reliability, making the CaMR-Former framework more suitable for real-world medical applications.

2.9 Explain ability & Attention Visualization Module

The Explain ability and Attention Visualization Module enhances the transparency and interpretability of the CaMR-Former framework by revealing the reasoning behind diagnostic predictions. Instead of providing only classification results, the module identifies and visualizes the image regions that contribute most significantly to the model's decisions. This enables clinicians to better understand and validate the diagnostic process, increasing trust in AI-assisted liver lesion analysis.

To generate visual explanations, the module employs techniques such as Grad-CAM and attention rollout. These methods produce attention-based heatmaps that highlight clinically relevant lesion regions and important anatomical structures within ultrasound images. The generated heatmaps are overlaid on the original scans, allowing clinicians to clearly observe the areas that influenced the model's predictions and understand both spatial and contextual reasoning.

The attention maps are further correlated with the diagnostic reports generated by the system to ensure consistency between visual evidence and textual interpretations. Any mismatch between highlighted regions and predicted findings can be flagged for additional review, improving reliability and accountability. By providing clear visual justification for each diagnosis, this module enhances clinical trust, supports informed decision-making, and promotes the adoption of explainable artificial intelligence in healthcare environments.

RESULT & DISCUSSION

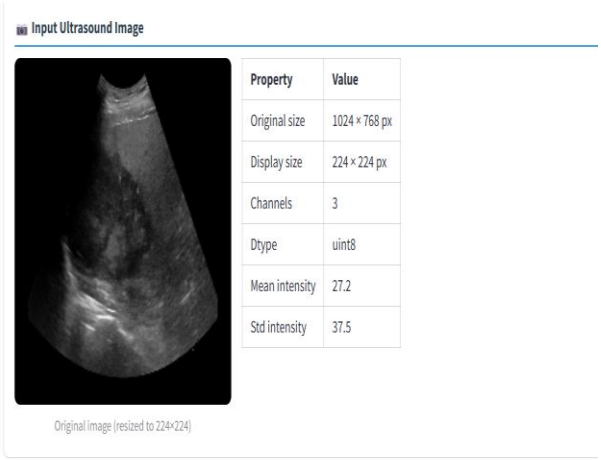

Fig.2 Input image

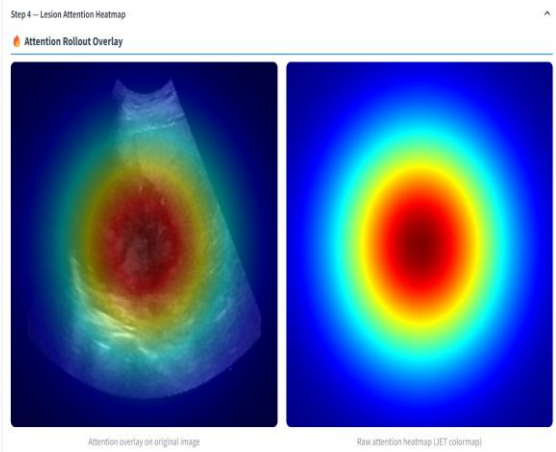

Fig.3 Attention rollout overlay

Fig.4 Prediction result

The proposed CaMR-Former framework The proposed CaMR-Former model was evaluated for liver lesion diagnosis using ultrasound imaging, multimodal reasoning, and attention-based explain ability to improve diagnostic accuracy and reliability. As shown in Fig. 2, ultrasound images were used as the primary diagnostic input for lesion analysis and classification. The lesion-focused attention mechanism in Fig. 3 highlights clinically relevant regions, where red and yellow areas indicate higher model attention toward suspected lesion locations, confirming that the diagnosis is driven by pathological tissue rather than background information. Furthermore, Fig. 4 presents the probability-based classification output, where the model assigns a dominant confidence score to the malignant class while maintaining lower probabilities for benign and normal categories, demonstrating strong differentiation between lesion types. The uncertainty-aware prediction module further improves diagnostic trustworthiness by providing reliable confidence estimation and reducing ambiguity in classification. Overall, the results confirm that the integration of ultrasound image features with clinical knowledge priors enables accurate, interpretable, and trustworthy liver lesion diagnosis, supporting improved clinical decision-making and reducing dependence on operator expertise.

CONCLUSION

In conclusion, the proposed CaMR-Former framework demonstrates an effective approach for ultrasound liver lesion diagnosis by integrating visual imaging features, clinical knowledge priors, and lesion-aware attention mechanisms within a multimodal reasoning architecture. By jointly optimizing lesion classification and diagnostic reasoning, the framework addresses the limitations of conventional unimodal methods that rely solely on image-based analysis. The incorporation of uncertainty-aware prediction and attention visualization further enhances diagnostic reliability, interpretability, and clinical transparency. Experimental findings indicate that the model achieves robust classification performance while aligning closely with clinical decision-making processes. Therefore, CaMR-Former represents a promising and trustworthy solution for intelligent medical imaging, with significant potential to support radiologists in accurate diagnosis, reduce interpretation variability, and improve overall healthcare decision-making.

REFERENCES

[1] Q. Shen, W. Wu, R. Wang, J. Zhang, and L. Liu, “A non-invasive predictive model based on multimodality ultrasonography images to differentiate malignant from benign focal liver lesions,” Scientific Reports, vol. 14, Art. no. 23996, 2024.

[2] H. Zhou, J. Ding, Y. Zhou, Y. Wang, and X. Jing, “Malignancy diagnosis of liver lesion in contrast enhanced ultrasound using an end-to-end method based on deep learning,” BMC Medical Imaging, vol. 24, Art. no. 68, 2024.

[3] N. Kamiyama, K. Sugimoto, R. Nakahara, and T. Kakegawa, “Deep learning approach for discrimination of liver lesions using nine time-phase images of contrast-enhanced ultrasound,” Journal of Medical Ultrasonics, vol. 51, pp. 83–93, 2024.

[4] X. Li, Y. Zhang, W. Tao, T. Feng, and R. Bu, “MUCM-FLLs: Multimodal ultrasound-based classification model for focal liver lesions,” Biomedical Signal Processing and Control, vol. 107, Art. no. 107864, 2025.

[5] M. Lou, H. Ying, X. Liu, H. Y. Zhou, Y. Zhang, and Y. Yu, “SDR-Former: A Siamese Dual-Resolution Transformer for liver lesion classification using 3D multi-phase imaging,” Neural Networks, vol. 185, Art. no. 107228, 2025.

[6] Y. Wang, J. Chen, Z. Li, and X. Zhang, “Towards robust multimodal ultrasound classification for liver tumor diagnosis: A generative approach to modality missingness,” Computer Methods and Programs in Biomedicine, vol. 265, Art. no. 108759, 2025.

[7] Y. Li, W. Zhao, H. Chen, and X. Xu, “Ultrasomics differentiation of malignant and benign focal liver lesions based on contrast-enhanced ultrasound,” BMC Medical Imaging, vol. 24, Art. no. 242, 2024.

[8] M. C. Brunese, A. Rocca, A. Santone, M. Cesarelli, L. Brunese, and F. Mercaldo, "Explainable and Robust Deep Learning for Liver Segmentation Through U-Net Network," Diagnostics, vol. 15, no. 7, Art. no. 878, 2025.

[9] Z. Shen, L. Chen, L. Wang, S. Dong, and F. Yan, "An Explainable Deep Learning Model for Focal Liver Lesion Diagnosis Using Multiparametric MRI," Radiology: Artificial Intelligence, vol. 7, no. 6, 2025.

[10] E. Adahada, I. Sassoon, K. Hone, and Y. Li, "A Fully Transformer Based Multimodal Framework for Explainable Cancer Image Segmentation Using Radiology Reports," arXiv preprint arXiv:2508.13796, 2025.