Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

ACMS–Net: A Lightweight Adaptive Multi-Scale Feature Learning with Cross-Scale Attention Network for Multi-Class Classification of Gallbladder Disease in Ultrasound Images

MD. RASHED[1], FAISAL IQBAL HIRON[1], MD. IMRAN HOSSAIN[1], MD. SARWAR HOSAIN[1], (Senior Member, IEEE), OHIDUJJAMAN[2], MD ZAHIDUL ISLAM[3], (Senior Member, IEEE), TETSUYA SHIMAMURA[4], (Senior Member, IEEE)

[1]Department of Information and Communication Engineering, Pabna University of Science and Technology, Pabna 6600, Bangladesh
[2]Department of Computer Science and Engineering, United International University, Dhaka 1212, Bangladesh
[3]Department of Information and Communication Technology, Islamic University, Kushtia 7003, Bangladesh
[4]Graduate School of Science and Engineering, Saitama University, 255 Shimo-Okubo, Sakura-ku, Saitama 338-8570, Japan

Corresponding authors: Md. Imran Hossain (e-mail: imran05ice@pust.ac.bd) and Md. Sarwar Hosain (e-mail: sarwar.ice@pust.ac.bd).

The research received no external funding. Article processing charges are supported by the Shimamura Laboratory, Saitama University, Japan

ABSTRACT Gallbladder (GB) disease covers a wide range of conditions, from gallstones and cholecystitis to adenomyomatosis and carcinoma. Early detection and classification is crucial, since GB carcinoma often show no clear symptoms until later stages. Recent deep learning models have pushed classification accuracy above 98% on this task but most of them lean on large backbones, capsule-routing designs, or transformer architectures that are simply too heavy to run in low-resource clinical settings, where US is often the only imaging option available. To close this research gaps, we propose ACMS-Net, A Lightweight Adaptive Multi-Scale Feature Learning with Cross-scale Attention Network for multi-class GB disease classification. It combines a four-branch Adaptive Multi-Scale Feature Extraction (AMSFE) block, a Cross-Scale Interaction Module (CSIM) built on query-key-value attention, an Adaptive Scale Attention mechanism that learns which scale matters most for a given input, and Squeeze-and-Excitation (SE) channel recalibration, all packed into a compact residual design. ACMS-Net reaches 99.91% accuracy, 99.98% specificity and an MCC of 99.89%, beating seventeen existing approaches while using only 0.85M parameters, 1.05 GFLOPs and 3.71 ms inference time per image (269.62 FPS). Five-fold cross-validation shows this performance holds steady across different data splits and an ablation study confirms that CSIM and Adaptive Scale Attention work together rather than overlapping in what they contribute. Overall, ACMS-Net provides a practical, real-time and interpretable option for gallbladder disease classification that is well suited to low-resource ultrasound workflows. Our code is publicly available at our GitHub repository: https://github.com/Rashed200613/ACMS-Net.

INDEX TERMS Gallbladder disease, ultrasound imaging, deep learning, multi-scale feature extraction, cross-scale attention, channel recalibration, lightweight network, medical image classification.

I. INTRODUCTION

GALLBLADDER (GB) is a small, pear-shaped organ situated beneath the liver, within the fossa formed between the liver's fourth and fifth segments; despite its simple, thin-walled structure, it performs the essential role of storing and concentrating bile produced by the liver to support fat digestion [1]–[3]. Gallbladder disease covers a wide range of different conditions, including gallstones, cholecystitis and its more severe membranous and gangrenous forms, gallbladder perforation, polyps and cholesterol crystals, adenomy-

omatosis, wall thickening and carcinoma. Each of these conditions looks different on an ultrasound (US) scan and telling them apart correctly is important for giving patients the right diagnosis and treatment [4], [5]. Some of these conditions are also linked to serious health outcomes. Around 86% of adenocarcinoma cases are found together with gallstones, showing how closely these two conditions are connected [6], [7]. On a global scale, GLOBOCAN 2022 reported more than 122,000 new cases of gallbladder carcinoma and close to 89,000 deaths in a single year. Because early-stage carcinoma usually shows no clear symptoms, it is often not caught until later stages, which lowers the chances of successful treatment [8], [9].

To make things more difficult, many of these conditions can look similar to each other on US. Wall thickening, for instance, can be caused by cholecystitis, adenomyomatosis, or even early carcinoma, so relying on wall thickening alone is not enough to tell these conditions apart [10]. This overlap makes multi-class classification a genuinely difficult task, since a model must learn to notice small, subtle differences rather than relying on a single obvious feature.

US is the most commonly used method for examining the gallbladder because it is safe, low-cost and gives results in real–time [11]. However, US images often contain noise and have lower clarity compared to other imaging methods like CT or MRI and the quality of the scan also depends heavily on the skill of the person operating the machine. This makes it harder to consistently detect subtle disease features, especially as the number of scans radiologists need to review continues to grow [12].

Deep learning has become a popular way to support radiologists by automatically classifying these US images. Some recent deep learning models such as GBCapsNet [13], GallNet [14], MSFE-GallNet-X [15], [16] and a hybrid model [17] have shown that it is possible to classify all nine disease categories with over 98% accuracy. Another approach, such as GB-YOLOv7 [18], introduced two attention mechanisms to focus on relevant image regions for detecting gallbladder cancer correctly. GBCHV-Trans [19] uses median filtering and CLAHE to enhance US image quality by reducing noise and improving contrast. They processed strip-based features by utilizing a self-attention-based GBCHV-Trans block to learn both global and local spatial dependencies for accurate cancer classification. [20]–[22], RadFormer- [23], Focusmae- [24] are employed advanced mechanisms such as feature fusion, attention, transformer learning, hierarchical classification, hyperparameter optimization and uncertainty estimation to improve diagnostic robustness and performance. These studies [25], [26] used CNN ensemble models and attention-based multiple instance learning (MIL) frameworks for accurate gallbladder cancer detection.

While these results are encouraging, they were achieved with high performance using large and complex models with a high number of parameters. Such models require more memory, more computing power and more time to run, which makes them difficult to use in hospitals with limited computing resources. A common situation in many developing regions where US scanning is often the only imaging tool available. Although high accuracy has already been shown to be possible for this multi-class classification task, the state-of-the-art models have no specific focus on making their models so lightweight and efficient while keeping accuracy high. A model that is both accurate and efficient would be more practical for real-world clinical use, especially in low-resource settings.

Motivated by these research gaps, our objectives are as follows: (1) To develop a lightweight and computationally efficient deep learning model (2) To capture both fine-grained local details and broader contextual information (3) To enhance feature representation through adaptive attention mechanisms (4) To improve feature selectivity while maintaining efficiency (5) To achieve high classification accuracy with reduced model complexity and (6) To facilitate reliable real-time computer-aided diagnosis in low-resource healthcare environments. Based on these objectives, this study proposes ACMS-Net, a lightweight adaptive cross-scale multi-scale attention network specifically designed for multi-class gallbladder disease classification from US images. This architecture combines efficient multi-scale feature extraction with cross-scale interaction and adaptive scale attention to effectively capture both fine-grained local structures and broader contextual information while maintaining low computational complexity. Furthermore, lightweight depthwise-separable convolutions, channel-wise feature recalibration and residual feature aggregation enhance the network to achieve an effective balance between diagnostic accuracy and computational efficiency. Consequently, ACMS-Net is well suited for real-time clinical deployment, particularly in low-resource healthcare settings where computational resources are limited and US is often the primary imaging modality.

The key contributions of this study are as follows:

- We propose **ACMS-Net**, a lightweight network that jointly integrates multi-scale feature extraction with cross-scale interaction and adaptive attention for efficient multi-class gallbladder disease classification from ultrasound images.
- We introduce a **Cross-Scale Interaction Module (CSIM)**, which enables bidirectional information exchange across multiple receptive fields through query-key-value attention, allowing each feature scale to be adaptively refined according to the global multi-scale context.
- We design an **Adaptive Scale Attention mechanism** that dynamically assigns input-dependent importance to different receptive-field scales, enabling adaptive selection of local, medium and wide-context features without manually predefined scale priorities.
- We demonstrate that combining **CSIM and adaptive scale attention within a single AMSFE** stage yields complementary scale recalibration at the inter-scale relational level (CSIM) and the global importance level (scale attention), which we validate through ablation on

gallbladder ultrasound classification.

II. RELATED WORKS

Recent progress in deep learning (DL) has led to the creation of many models for solving specific problems of classification accuracy and computational efficiency. In this section, some of the existing studies on gallbladder (GB) abnormality classification and identification using different methodological approaches are discussed.

A. TRANSFER LEARNING-BASED CNN ARCHITECTURES

A dominant line of work adapts pre-trained Convolutional Neural Network (CNN) backbones for GB disease classification. This line of research [17] is exemplified by a method that extracted VGG16 features and classified them using a Grid-Search-optimized SVM, reaching 99.35% accuracy on multi-class GB ultrasound classification. This paper [27] employs four transfer-learning CNNs combined with bilateral filtering and active contour segmentation to classify ultrasound images, demonstrating superior performance in identifying multiple diseases of the gallbladder. Another study [21] introduced a method that combined Inception-based transfer learning with Gorilla-Troops hyperparameter optimization and a BiGRU classifier, resulting in an accuracy of 96.29%. In contrast, Elbedwehy et al. [28] fused EfficientViT and AlexNet through an MLP block optimized with the Lion optimizer, achieving an impressive accuracy of 99.86% and providing interpretability through Grad-CAM. Extending this direction to a distributed clinical setting, a federated VGG16 model can outperform its locally trained counterpart by achieving 99% accuracy, underscoring the value of privacy-preserving multi-institutional training [29]. A ResNet50 [30] model applied to CT patches to differentiate GB cancer from xanthogranulomatous cholecystitis achieved 98.5% accuracy, indicating that transfer-learning pipelines generalize across imaging modalities but remain fundamentally reliant on off-the-shelf backbones rather than task-specific architectural design.

B. CAPSULE NETWORKS AND ATTENTION–RECURRENT HYBRID MODELS

Capsule based architectures have been explored to better preserve the spatial and part-whole relationships lost in standard pooling. GBCapsNet [13] achieved an impressive accuracy of 99.91% by combining convolutional feature extraction with dynamic routing and temperature-scaling calibration. In a related approach, an SE-CapsNet [16] feature extraction module was integrated with a CNN-BiLSTM classifier, resulting in a slightly lower accuracy of 99.09%. However, this method demonstrated the advantages of combining channel attention with the sequence modeling of capsule outputs. This suggests that it is possible to enhance architectural interpretability and temporal feature modeling without significantly compromising performance.

C. MULTI-SCALE AND SELF-ATTENTION-BASED FEATURE EXTRACTION APPROACHES

MSFE-GallNet-X [15] employed parallel convolutional kernels for multi-scale feature extraction, coupled with Grad-CAM and LIME interpretability, achieving 99.63% accuracy. A self-attention-guided residual network with dilated multi-scale convolutions [31] achieved 99.17% accuracy and 98.94% recall.

D. TRANSFORMER-BASED, ANATOMY-AWARE AND MULTIPLE-INSTANCE LEARNING FRAMEWORKS

Another line of work is global-context modeling. GBCHV [19] is an anatomy-aware transformer that utilizes horizontal-vertical strip embeddings applied after median filtering and CLAHE (Contrast Limited Adaptive Histogram Equalization) enhancement, achieving an impressive accuracy of 96.21%. RadFormer [23] integrates a ResNet-50 global branch with a BagNet-based local branch via a transformer, surpassing expert radiologists with an accuracy of 92.1%, while still maintaining interpretability through bag-of-features explanations. A gated-attention multiple-instance learning (MIL) framework with an EfficientNet-B4 backbone [26] was validated across multiple centers, achieving an area under the curve (AUC) of 0.883 on the internal test set. Additionally, an ensemble of ResNet152, InceptionV3 and DenseNet161, combined with clinical metadata (such as age and polyp size) [25], was employed for polyp classification, attaining an accuracy of 87.61%. Collectively, these studies indicate that global-context aggregation—whether through transformers or attention-based MIL—is crucial for robustness across diverse, multi-center data, although this often comes with a significantly higher parameter count.

E. DETECTION-BASED, TEMPORAL AND SELF-SUPERVISED LEARNING APPROACHES

Several studies have redefined gallbladder (GB) classification as a problem of detection or sequence modeling. A YOLOv7-based detector integrating Normalization-based Attention Modules (NAM) and Global Attention Modules (GAM) [18] was proposed for gallbladder cancer detection, achieving 91.3% recall and outperforming YOLOv7, YOLOv8, YOLOv9 and YOLOv11 existings while using fewer parameters. RetinaNet [22] was employed for gallbladder region detection, followed by hierarchical classification using GBCNet with added uncertainty calibration to improve prediction reliability, achieving a classification accuracy of 92.62%. . Temporal cues have also been exploited: Focus-MAE [24], a region-guided masked-autoencoder framework for video-based GB cancer detection, achieved 96.4% accuracy, while an InceptionV3-GRU [32] model for cine ultrasound sequence modeling achieved 91% and 82% accuracy on two binary classification tasks. In a data-efficient direction, ASGBC [20] combined active and self-supervised learning with a multi-scale high-order pooling module, matching competitive performance using only 35% of labeled training data and achieving 88.4% accuracy with 91.2% sensitivity.

Table 1 represent the Comparative summary of recent deep learning-based studies on gallbladder disease classification and detection, highlighting model architectures, datasets, key performance and limitations.

F. SYNTHESIS OF RESEARCH GAPS

Most existing gallbladder classification models applied in these studies are based on multi-scale extraction, attention or channel recalibration individually, but they cannot jointly model the cross-scale interaction and adaptive, sample-dependent scale importance. Furthermore, the most accurate methods tend to depend on large backbones or additional sequence-modeling components to achieve their accuracy, leading to large model sizes that are not practical to deploy in low-resource clinical settings where ultrasound is often the primary imaging modality. This leaves a clear gap for a unified and lightweight architecture that combines multi-scale feature extraction, cross-scale interaction and adaptive scale attention into a single efficient design-motivated by the proposed ACMS-Net.

III. PROPOSED METHODOLOGY

This section presents the proposed Adaptive Cross-Scale Multi-Scale Attention Network (ACMS–Net) for multi-class gallbladder disease classification from US images. Gallbladder pathologies exhibit substantial variability in lesion size, wall echogenicity and spatial extent, ranging from focal abnormalities such as polyps and gallstones to diffuse conditions such as chronic cholecystitis and carcinoma. Effectively distinguishing among these categories therefore requires a model capable of jointly reasoning over both fine-grained local textures and broader contextual patterns and of adaptively emphasizing whichever scale of information is most discriminative for a given input. ACMS-Net is designed around this requirement through four core components: (i) a lightweight convolutional stem for low-level feature extraction, (ii) an Adaptive Multi-Scale Feature Extraction (AMSFE) block that captures features at four complementary receptive-field scales using point-wise, standard, depthwise-separable and dilated depthwise convolutions, (iii) a Cross-Scale Interaction Module (CSIM) and an Adaptive Scale Attention mechanism that jointly recalibrate these scale-specific features based on their mutual contextual relevance and their sample-dependent importance and (iv) a Squeeze-and-Excitation (SE) block for fine-grained channel-wise recalibration. Two such AMSFE stages are stacked, connected through a lightweight transition (bridge) block and followed by a convolutional refinement block and a fully connected classification head. The overall architecture is illustrated in Figure 1 and the mathematical formulation of each component is detailed in the subsequent subsections, following the problem formulation and end-to-end network composition given below.

A. CONVOLUTIONAL STEM

The convolutional stem $f_{Stem}(\cdot)$ performs initial low-level feature extraction from the raw US image, converting the raw pixel intensities into an early feature representation while suppressing high-frequency acquisition noise common in B–mode US imaging. It consists of two successive 3×3 convolutional layers, each followed by batch normalization and ReLU activation and a final 2×2 max-pooling operation for spatial down-sampling.

$$\mathbf{Z}_1 = \mathrm{ReLU}\big(\mathrm{BN}(\mathrm{Conv}_{3\times3}^{3\to32}(\mathbf{X}))\big), \tag{1}$$

$$\mathbf{Z}_2 = \mathrm{ReLU}\big(\mathrm{BN}(\mathrm{Conv}_{3\times3}^{32\to32}(\mathbf{Z}_1))\big), \tag{2}$$

$$\mathbf{F}_0 = \mathrm{MaxPool}_{2\times2}(\mathbf{Z}_2), \tag{3}$$

Here, $\mathbf{X} \in \mathbb{R}^{3\times224\times224}$ is the input RGB US image ($H = W = 224$). $\mathrm{Conv}_{3\times3}^{3\to32}(\cdot)$ and $\mathrm{Conv}_{3\times3}^{32\to32}(\cdot)$ denote 3×3 convolutions (stride 1, padding 1, no bias) with weights $W_{s_1} \in \mathbb{R}^{32\times3\times3\times3}$ and $W_{s_2} \in \mathbb{R}^{32\times32\times3\times3}$, respectively, each followed by batch normalization $\mathrm{BN}(\cdot)$ and $\mathrm{ReLU}(\cdot) = \max(0, \cdot)$. $\mathrm{MaxPool}_{2\times2}(\cdot)$ applies stride-2 non-overlapping max-pooling. The stem output is $\mathbf{F}_0 \in \mathbb{R}^{32\times112\times112}$.

B. ADAPTIVE MULTI-SCALE FEATURE EXTRACTION (AMSFE) BLOCK

To capture disease-relevant patterns occurring at different spatial extents (e.g., localized wall thickening versus diffuse echogenic changes), the AMSFE block extracts features through four parallel branches with progressively larger effective receptive fields: point-wise, medium-scale, large-scale depthwise-separable and dilated depthwise convolutions.

Given an input feature map $\mathbf{F} \in \mathbb{R}^{C_{in}\times H\times W}$, where C_{in} denotes the number of input channels and $H \times W$ the spatial resolution at the current stage, each branch produces C output channels ($C = 42$):

$$\mathbf{S}_1 = \mathrm{ReLU}\big(\mathrm{BN}(\mathrm{Conv}_{1\times1}(\mathbf{F}))\big), \tag{4}$$

$$\mathbf{S}_3 = \mathrm{ReLU}\big(\mathrm{BN}(\mathrm{Conv}_{3\times3}(\mathbf{F}))\big), \tag{5}$$

$$\mathbf{S}_5 = \mathrm{ReLU}\Big(\mathrm{BN}\big(\mathrm{Conv}_{1\times1}(\mathrm{DWConv}_{5\times5}(\mathbf{F}))\big)\Big), \tag{6}$$

$$\mathbf{S}_7 = \mathrm{ReLU}\Big(\mathrm{BN}\big(\mathrm{Conv}_{1\times1}(\mathrm{DWConv}_{7\times7}^{d=2}(\mathbf{F}))\big)\Big), \tag{7}$$

Eq. (4): $\mathrm{Conv}_{1\times1}(\cdot)$ is a point-wise convolution with weight $W_{b_1} \in \mathbb{R}^{C\times C_{in}\times1\times1}$; $\mathbf{S}_1 \in \mathbb{R}^{C\times H\times W}$ is the branch output. Eq. (5): $\mathrm{Conv}_{3\times3}(\cdot)$ is a standard convolution with weight $W_{b_3} \in \mathbb{R}^{C\times C_{in}\times3\times3}$ and padding 1; $\mathbf{S}_3 \in \mathbb{R}^{C\times H\times W}$ is the branch output. Eq. (6): $\mathrm{DWConv}_{5\times5}(\cdot)$ is a depthwise convolution with weight $W_{d_5} \in \mathbb{R}^{C_{in}\times1\times5\times5}$, padding 2, groups $= C_{in}$; the subsequent $\mathrm{Conv}_{1\times1}(\cdot)$ is a point-wise projection with weight $W_{p_5} \in \mathbb{R}^{C\times C_{in}\times1\times1}$; $\mathbf{S}_5 \in \mathbb{R}^{C\times H\times W}$ is the branch output. Eq. (7): $\mathrm{DWConv}_{7\times7}^{d=2}(\cdot)$ is a dilated depthwise convolution with weight $W_{d_7} \in \mathbb{R}^{C_{in}\times1\times7\times7}$, dilation $d = 2$, padding 6; the subsequent $\mathrm{Conv}_{1\times1}(\cdot)$ is a point-wise projection with weight $W_{p_7} \in \mathbb{R}^{C\times C_{in}\times1\times1}$; $\mathbf{S}_7 \in \mathbb{R}^{C\times H\times W}$ is the branch output.

TABLE 1. Comparative Summary of Recent Deep Learning-Based Studies on Gallbladder Disease Classification and Detection, Highlighting Model Architectures, Datasets, Key Performance, and Limitations

Ref.	Approach	Dataset	Key Results	Limitations
[17]	VGG16 feature extraction, Grid-Search-optimized SVM	UIdataGB	99.35% accuracy	Relies on a handcrafted ML classifier stacked on CNN features rather than end-to-end learning
[27]	Transfer-learning backbones after NLM enhancement, ROI segmentation	GBCU Dataset	98.35% accuracy (best: MobileNet)	Gains driven mainly by preprocessing (NLM, ROI), not architectural design
[21]	Inception-based TL, gorilla-troops hyperparameter optimization, BiGRU	GBCU Dataset	96.29% accuracy	Added optimization/BiGRU increases complexity without multi-scale feature modeling
[28]	EfficientViT, AlexNet fused via MLP block, Lion optimizer	GBCU Dataset, curated UIdataGB, AURORA-GB	99.86% accuracy, Grad-CAM interpretability	Fusion of two heavyweight backbones increases model size/complexity
[29]	Federated learning across multiple institutions with VGG16	UIdataGB	99% accuracy	Federated setup adds communication overhead; no scale/attention innovation
[13]	Capsule network with dynamic routing, temperature-scaling calibration	UIdataGB	99.91% accuracy	Capsule routing is computationally costly; no explicit multi-scale extraction
[16]	SE-CapsNet feature extraction, BiLSTM sequence classifier	UIdataGB	99.09% accuracy	Only channel-wise attention, no spatial multi-scale extraction; BiLSTM adds cost
[15]	Multi-scale CNN with parallel kernels, Grad-CAM and LIME	UIdataGB	99.63% accuracy, 1.91M params	Fixed scale weighting; no cross-scale interaction
[31]	Self-attention-guided residual deep neural network with multi-scale dilated feature extraction	UIdataGB	99.17% accuracy, 98.94% recall	Global self-attention only; no per-scale query-key-value interaction
[19]	Anatomy-aware transformer with horizontal-vertical strip embeddings, median filter, CLAHE	GBCU Dataset	96.21% accuracy	Transformer-based design adds computational cost
[23]	ResNet-50 global branch, BagNet local branch, transformer	GBCU Dataset	92.1% accuracy (beat radiologists)	Complex dual-branch architecture; large parameter count
[26]	EfficientNet-B4 backbone with gated-attention MIL	AURORA-GB Dataset, GBCU Dataset, UIdataGB Dataset	AUC 0.883	MIL framework adds complexity; mainly validated on internal test set
[18]	YOLOv7 detector with NAM, GAM attention modules	GBCU Dataset	91.3% recall, 93.9% mAP, 24.34M params	Detection-based rather than classification; still computationally heavy
[22]	RetinaNet detection, hierarchical GBCNet classification, uncertainty calibration	GB ultrasound	92.62% accuracy	Two-stage detection-classification pipeline adds complexity
[24]	Region-guided masked-autoencoder, video-based detection	GB ultrasound videos	96.4% accuracy	Video/temporal modeling increases computational demand
[32]	InceptionV3, GRU for cine ultrasound sequences	Cine ultrasound sequences	91% & 82% accuracy (2 tasks)	Lower accuracy on one task; sequence modeling adds overhead
[20]	Active, self-supervised learning with multi-scale high-order pooling	GB Ultrasound (35% labeled)	88.4% accuracy, 91.2% sensitivity	Lower accuracy than fully supervised methods

In all four equations, $\mathrm{BN}(\cdot)$ and $\mathrm{ReLU}(\cdot)$ denote batch normalization and the rectified linear unit, as defined previously.

The four branch outputs are concatenated along the channel dimension, as shown in Eq. (8):

$$\mathbf{F}_{AMSFE} = \left[\mathbf{S}_1 \,\|\, \mathbf{S}_3 \,\|\, \mathbf{S}_5 \,\|\, \mathbf{S}_7\right] \in \mathbb{R}^{4C\times H\times W}. \quad (8)$$

In Eq. (8), $[\cdot \,\|\, \cdot]$ denotes channel-wise concatenation and $\mathbf{F}_{AMSFE} \in \mathbb{R}^{4C\times H\times W}$ (with $4C = 168$) is the resulting multi-scale feature map, in which each of the four scale-specific representations occupies a distinct, contiguous block of $C = 42$ channels: $\mathbf{F}_{AMSFE}[0:C] = \mathbf{S}_1$, $\mathbf{F}_{AMSFE}[C:2C] = \mathbf{S}_3$, $\mathbf{F}_{AMSFE}[2C:3C] = \mathbf{S}_5$, $\mathbf{F}_{AMSFE}[3C:4C] = \mathbf{S}_7$. This channel-block layout subsequently allows the CSIM and scale-attention modules to address each scale independently via simple channel slicing.

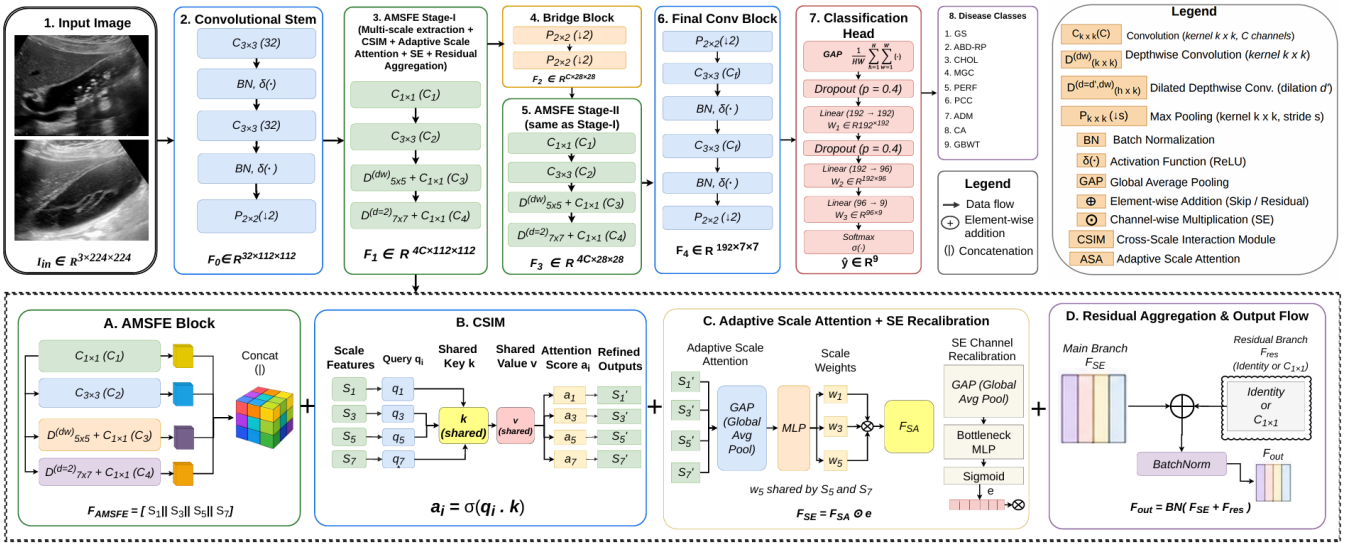

FIGURE 1. Overall architecture of the proposed ACMS–Net. The network begins with a convolutional stem for low-level feature extraction, followed by two Adaptive Multi-Scale Feature Extraction (AMSFE) stages connected through a transition (bridge) block. Each AMSFE stage integrates four-branch multi-scale feature extraction, a Cross-Scale Interaction Module (CSIM), Adaptive Scale Attention, and Squeeze-and-Excitation (SE) channel recalibration, together with a residual connection. The final convolutional block, global average pooling, and fully connected classifier produce the nine-class disease prediction.

C. CROSS-SCALE INTERACTION MODULE (CSIM)

While the AMSFE block extracts scale-specific features independently, disease patterns often manifest jointly across scales (e.g., a locally textured lesion embedded in a diffusely altered background). The CSIM enables each scale-specific feature map to be recalibrated using contextual information aggregated from all scales jointly.

For each scale $i \in \{1, 3, 5, 7\}$, a query descriptor is obtained by global average pooling the scale-specific feature map followed by a linear projection with reduction ratio $r = 4$:

$$\mathbf{q}_i = \mathrm{ReLU}\big(W_{q_i}\,\mathrm{GAP}(\mathbf{S}_i)\big) \in \mathbb{R}^{C/r}. \tag{9}$$

Here, $\mathrm{GAP}(\mathbf{S}_i) = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathbf{S}_i[:,h,w] \in \mathbb{R}^C$ is global average pooling, which squeezes the spatial dimensions of the scale-specific feature map $\mathbf{S}_i$ into a single C-dimensional channel descriptor summarizing its overall activation pattern. $W_{q_i} \in \mathbb{R}^{\lfloor C/r \rfloor \times C}$ is a scale-specific, learnable linear projection (with $r = 4$, so $C/r = 42/4 = 10.5 \rightarrow$ implemented as $\lfloor C/r \rfloor = 10$ channels) that maps this descriptor into a lower-dimensional query space; a distinct W_{q_i} exists for each of the four scales $i \in \{1, 3, 5, 7\}$, allowing each scale to learn its own query representation. $\mathbf{q}_i \in \mathbb{R}^{C/r}$ is the resulting query vector for scale i, intuitively representing "what this scale is looking for" in the joint multi-scale context.

A shared key descriptor and a shared value descriptor are computed from the full concatenated multi-scale feature map $\mathbf{F}_{AMSFE}$:

$$\mathbf{k} = \mathrm{ReLU}\big(W_k\,\mathrm{GAP}(\mathbf{F}_{AMSFE})\big) \in \mathbb{R}^{C/r}, \tag{10}$$
$$\mathbf{v} = \sigma\big(W_v\,\mathrm{GAP}(\mathbf{F}_{AMSFE})\big) \in \mathbb{R}^{C}, \tag{11}$$

In Eq. (10), $\mathrm{GAP}(\mathbf{F}_{AMSFE}) \in \mathbb{R}^{4C}$ squeezes the entire multi-scale feature map (all four scales jointly) into a single $4C$-dimensional descriptor. $W_k \in \mathbb{R}^{(C/r)\times4C}$ projects this joint descriptor into the same query space dimensionality, producing the shared key $\mathbf{k} \in \mathbb{R}^{C/r}$, which represents the aggregate context against which every scale's query is compared. In Eq. (11), $W_v \in \mathbb{R}^{C\times4C}$ projects the same joint descriptor into a C-dimensional space and $\sigma(\cdot) = 1/(1+e^{-(\cdot)})$ is the element-wise sigmoid function, bounding the shared value vector $\mathbf{v} \in \mathbb{R}^C$ to $(0, 1)$ so that it can act as a channel-wise gating signal.

A scalar cross-scale attention score for scale i is computed as a sigmoid-normalized dot product between its query and the shared key:

$$a_i = \sigma\big(\mathbf{q}_i \cdot \mathbf{k}\big) \in \mathbb{R}. \tag{12}$$

Here $\mathbf{q}_i \cdot \mathbf{k} = \sum_{j=1}^{C/r} q_{i,j}k_j$ is the inner product between the scale-i query and the shared key, a scalar measuring the compatibility (alignment) between scale i's local descriptor and the global multi-scale context; passing it through $\sigma(\cdot)$ produces the attention score $a_i \in (0, 1)$, which is broadcast (per sample in the batch) and does not vary spatially.

The shared value vector is gated by this attention score and used to recalibrate the corresponding scale-specific feature map channel-wise:

$$\mathbf{S}'_i = \mathbf{S}_i \odot \big(a_i \cdot \mathbf{v}\big), \qquad i \in \{1, 3, 5, 7\}, \tag{13}$$

Here, $a_i \cdot \mathbf{v} \in \mathbb{R}^C$ scales the shared value vector by the scale-specific attention score a_i, yielding a channel-wise gate specific to scale i; $\odot$ denotes channel-wise (broadcast) multiplication, in which each of the C gate values multiplies every spatial location (h, w) of the corresponding channel in $\mathbf{S}_i$, without altering the spatial resolution. $\mathbf{S}'_i \in \mathbb{R}^{C\times H\times W}$ is

the resulting cross-scale-recalibrated feature map for scale i: channels of $\mathbf{S}_i$ that are more relevant given the global multi-scale context (as measured by a_i and $\mathbf{v}$) are amplified, while less relevant channels are suppressed.

The recalibrated scale features are concatenated to form the CSIM output, preserving the input shape:

$$\mathbf{F}_{CSIM} = \left[\mathbf{S}'_1 \,\|\, \mathbf{S}'_3 \,\|\, \mathbf{S}'_5 \,\|\, \mathbf{S}'_7\right] \in \mathbb{R}^{4C\times H\times W}. \tag{14}$$

$\mathbf{F}_{CSIM} \in \mathbb{R}^{4C\times H\times W}$ has identical shape to $\mathbf{F}_{AMSFE}$, since CSIM only re-weights channels without changing spatial resolution or channel count; it is passed forward to the scale attention module.

D. ADAPTIVE SCALE ATTENTION

Not all scales contribute equally to every disease category; some conditions are best characterized by local texture, while others require wider contextual cues. The scale attention module learns a soft, input-dependent weighting over three scale groups: point-wise (w_1), medium-scale (w_3) and a merged large-scale/wide-context group (w_5, shared between the 5×5 and dilated 7×7 branches).

$$\mathbf{h} = \mathrm{ReLU}\big(W_1\,\mathrm{GAP}(\mathbf{F}_{CSIM})\big), \tag{15}$$

$$[\alpha_1, \alpha_3, \alpha_5] = W_2\,\mathbf{h}, \tag{16}$$

$$[w_1, w_3, w_5] = \mathrm{Softmax}([\alpha_1, \alpha_3, \alpha_5]), \tag{17}$$

In Eq. (15), $\mathrm{GAP}(\mathbf{F}_{CSIM}) \in \mathbb{R}^{4C}$ squeezes the CSIM output into a $4C$-dimensional descriptor summarizing the joint multi-scale activation statistics of the current sample. $W_1 \in \mathbb{R}^{(4C/4)\times 4C}$ is a learnable projection reducing this descriptor to a bottleneck dimensionality of $4C/4 = C = 42$ and $\mathbf{h} \in \mathbb{R}^{4C/4}$ is the resulting hidden representation after ReLU. In Eq. (16), $W_2 \in \mathbb{R}^{3\times(4C/4)}$ projects $\mathbf{h}$ to three raw scale-importance logits $\alpha_1, \alpha_3, \alpha_5 \in \mathbb{R}$, one per scale group. In Eq. (17), $\mathrm{Softmax}([\alpha_1, \alpha_3, \alpha_5])_k = \exp(\alpha_k)/\sum_j \exp(\alpha_j)$ normalizes these logits into non-negative weights that sum to one, yielding the three scale-importance weights $w_1, w_3, w_5 \in (0, 1)$, such that

$$w_1 + w_3 + w_5 = 1. \tag{18}$$

This constraint guarantees that the three weights form a valid convex combination, so that the subsequent rescaling in Eq. (19) redistributes, rather than amplifies or attenuates, the overall feature energy across scale groups.

The scale-weighted feature map is then obtained by splitting $\mathbf{F}_{CSIM}$ back into its four constituent branches and rescaling each by its corresponding weight, with the 5×5 and dilated 7×7 branches sharing w_5:

$$\mathbf{F}_{SA} = \left[\, w_1\mathbf{S}'_1 \,\|\, w_3\mathbf{S}'_3 \,\|\, w_5\mathbf{S}'_5 \,\|\, w_5\mathbf{S}'_7 \,\right] \in \mathbb{R}^{4C\times H\times W}. \tag{19}$$

Here, each scalar weight w_1, w_3, w_5 multiplies every channel and every spatial location of its corresponding scale-specific block ($\mathbf{S}'_1$, $\mathbf{S}'_3$ and jointly $\mathbf{S}'_5$, $\mathbf{S}'_7$, respectively), so that scale groups deemed less discriminative for the current input are proportionally down-weighted before being re-concatenated into $\mathbf{F}_{SA} \in \mathbb{R}^{4C\times H\times W}$, which retains the same shape as $\mathbf{F}_{CSIM}$.

E. SQUEEZE-AND-EXCITATION (SE) CHANNEL RECALIBRATION

Following scale-level weighting, a Squeeze-and-Excitation block performs fine-grained, per-channel recalibration across all $4C$ channels jointly.

$$\mathbf{z} = \mathrm{GAP}(\mathbf{F}_{SA}) \in \mathbb{R}^{4C}, \tag{20}$$

$$\mathbf{e} = \sigma\big(W_{e2}\,\mathrm{ReLU}(W_{e1}\mathbf{z})\big) \in \mathbb{R}^{4C}, \tag{21}$$

$$\mathbf{F}_{SE} = \mathbf{F}_{SA} \odot \mathbf{e}, \tag{22}$$

In Eq. (20), the squeeze operation $\mathrm{GAP}(\mathbf{F}_{SA})$ compresses the spatial extent of $\mathbf{F}_{SA}$ into a single $4C$-dimensional channel descriptor $\mathbf{z} \in \mathbb{R}^{4C}$, summarizing the global activation level of each of the 168 channels. In Eq. (21), the excitation operation applies a bottlenecked two-layer transformation: $W_{e1} \in \mathbb{R}^{\lfloor 4C/16\rfloor\times 4C}$ reduces the channel descriptor to a compact bottleneck of dimensionality $\lfloor 4C/16\rfloor = \lfloor 168/16\rfloor = 10$ (reduction ratio 16, implemented as $4C \,//\, 16$ channels, consistent with the integer-division handling used for the query projection in the Cross-Scale Interaction Module, Section 3.3), $\mathrm{ReLU}(\cdot)$ introduces nonlinearity, $W_{e2} \in \mathbb{R}^{4C\times\lfloor 4C/16\rfloor}$ projects it back to the original $4C$ dimensionality and $\sigma(\cdot)$ bounds the result to $(0, 1)$, producing the channel-wise excitation vector $\mathbf{e} \in \mathbb{R}^{4C}$ that encodes the relative importance of each channel. In Eq. (22), $\odot$ again denotes channel-wise broadcast multiplication, so that $\mathbf{F}_{SE} \in \mathbb{R}^{4C\times H\times W}$ is obtained by scaling every spatial location of each channel of $\mathbf{F}_{SA}$ by its corresponding excitation value in $\mathbf{e}$, adaptively emphasizing informative channels and suppressing less useful ones irrespective of which scale branch they originated from.

F. AMSFE STAGE (RESIDUAL AGGREGATION)

Each complete AMSFE stage combines the operations above with a residual connection, allowing the network to retain the original input representation alongside the newly extracted multi-scale features and improving gradient flow during backpropagation.

$$\mathbf{F}_{res} = \begin{cases} \mathbf{F}, & \text{if } C_{in} = 4C, \\ \mathrm{Conv}_{1\times 1}^{C_{in}\to 4C}(\mathbf{F}), & \text{otherwise}, \end{cases} \tag{23}$$

Here, $\mathbf{F}_{res} \in \mathbb{R}^{4C\times H\times W}$ is the residual (identity or projected) branch. When the number of input channels C_{in} already equals the stage output width $4C$, the residual branch is the identity mapping $\mathbf{F}_{res} = \mathbf{F}$, requiring no additional parameters. Otherwise (e.g., at the first AMSFE stage, where $C_{in} = 32 \neq 4C = 168$), $\mathrm{Conv}_{1\times 1}^{C_{in}\to 4C}(\cdot)$ is a learnable 1×1 convolution (weight $W_{res} \in \mathbb{R}^{4C\times C_{in}\times 1\times 1}$, no bias) that linearly projects $\mathbf{F}$ into the $4C$-dimensional space so that it can be added element-wise to the main branch output.

$$f_{AMSFE}(\mathbf{F}) = \mathrm{BN}\big(\mathbf{F}_{SE} + \mathbf{F}_{res}\big) \in \mathbb{R}^{4C\times H\times W}, \quad (24)$$

where $\mathbf{F}_{SE}$ is obtained by sequentially applying Eqs. (8)–(22) to $\mathbf{F}$. The element-wise sum $\mathbf{F}_{SE}+\mathbf{F}_{res}$ fuses the multi-scale, cross-scale-, scale- and channel-recalibrated features with the (identity- or projection-based) input representation and a final batch normalization $\mathrm{BN}(\cdot)$ stabilizes the combined activations. With $C = 42$, each AMSFE stage outputs $4C = 168$ channels.

Applying this stage to the stem output yields the first-stage representation:

$$\mathbf{F}_1 = f^{(1)}_{AMSFE}(\mathbf{F}_0) \in \mathbb{R}^{168\times 112\times 112}. \quad (25)$$

Here $f^{(1)}_{AMSFE}(\cdot)$ denotes the first AMSFE stage instantiated with $C_{in} = 32$ (matching the stem output channel count) and branch width $C = 42$; since $C_{in} = 32 \neq 168$, its residual branch uses the 1×1 projection case of Eq. (23). $\mathbf{F}_1 \in \mathbb{R}^{168\times 112\times 112}$ preserves the spatial resolution of $\mathbf{F}_0$, since every operation within the AMSFE stage is spatial-resolution-preserving by construction.

G. TRANSITION (BRIDGE) BLOCK

Between the two AMSFE stages, a lightweight transition block reduces spatial resolution while introducing no additional learnable parameters, keeping the network compact prior to the second stage of multi-scale extraction.

$$\hat{\mathbf{F}}_1 = \mathrm{MaxPool}_{2\times 2}(\mathbf{F}_1) \in \mathbb{R}^{168\times 56\times 56}, \quad (26)$$

$$\mathbf{F}_2 = f_{Bridge}(\hat{\mathbf{F}}_1) = \mathrm{MaxPool}_{2\times 2}(\hat{\mathbf{F}}_1) \in \mathbb{R}^{168\times 28\times 28}. \quad (27)$$

In Eq. (26), $\mathrm{MaxPool}_{2\times 2}(\cdot)$ first halves the spatial resolution of $\mathbf{F}_1$ (from 112×112 to 56×56) without altering its 168 channels, producing the intermediate map $\hat{\mathbf{F}}_1$. A second $\mathrm{MaxPool}_{2\times 2}(\cdot)$ in Eq. (27) then halves the resolution again (from 56×56 to 28×28), yielding the bridge output $\mathbf{F}_2 \in \mathbb{R}^{168\times 28\times 28}$, which serves as the input to the second AMSFE stage. Unlike a conventional transition block, no intermediate 1×1 convolution is applied here: since the channel count is unchanged across the bridge ($168 \rightarrow 168$) and the subsequent AMSFE stage already performs extensive channel-wise re-mixing through its four parallel branches, CSIM and SE recalibration, an additional point-wise projection at this stage was found to add parameters without a corresponding benefit to representational capacity. The transition block is therefore reduced to pure spatial downsampling, denoted by the composite function $f_{Bridge}(\cdot)$.

H. SECOND AMSFE STAGE

The second AMSFE stage re-applies the full sequence of Eqs. (8)–(24) to $\mathbf{F}_2$, with input channels $C_{in} = 168$ and the same per-branch width $C = 42$:

$$\mathbf{F}_3 = f^{(2)}_{AMSFE}(\mathbf{F}_2) \in \mathbb{R}^{168\times 28\times 28}. \quad (28)$$

Here, $f^{(2)}_{AMSFE}(\cdot)$ denotes the second AMSFE stage, structurally identical in design to $f^{(1)}_{AMSFE}(\cdot)$ but with its own independently learned parameters (branch convolutions, CSIM projections, scale-attention weights, SE weights). Since its input channel count $C_{in} = 168$ already equals the stage output width $4C = 168$, its residual branch uses the identity case of Eq. (23), i.e., $\mathbf{F}_{res} = \mathbf{F}_2$, requiring no additional 1×1 projection parameters. The output $\mathbf{F}_3 \in \mathbb{R}^{168\times 28\times 28}$ preserves the spatial resolution of $\mathbf{F}_2$.

I. FINAL CONVOLUTIONAL BLOCK

After the second AMSFE stage, a lightweight convolutional block widens the channel dimension and further condenses spatial information prior to classification, forming higher-level, task-specific feature combinations while keeping parameter growth minimal.

$$\hat{\mathbf{F}}_3 = \mathrm{MaxPool}_{2\times 2}(\mathbf{F}_3) \in \mathbb{R}^{168\times 14\times 14}, \quad (29)$$

$$\mathbf{Z}_3 = \mathrm{ReLU}\big(\mathrm{BN}(\mathrm{Conv}^{168\rightarrow 192}_{3\times 3}(\hat{\mathbf{F}}_3))\big), \quad (30)$$

$$\mathbf{Z}_4 = \mathrm{ReLU}\big(\mathrm{BN}(\mathrm{Conv}^{192\rightarrow 192}_{3\times 3}(\mathbf{Z}_3))\big), \quad (31)$$

$$\mathbf{F}_4 = \mathrm{MaxPool}_{2\times 2}(\mathbf{Z}_4) \in \mathbb{R}^{192\times 7\times 7}. \quad (32)$$

In Eq. (29), $\mathrm{MaxPool}_{2\times 2}(\cdot)$ halves the spatial resolution of $\mathbf{F}_3$ from 28×28 to 14×14, yielding $\hat{\mathbf{F}}_3 \in \mathbb{R}^{168\times 14\times 14}$. In Eq. (30), $\mathrm{Conv}^{168\rightarrow 192}_{3\times 3}(\cdot)$ (weight $W_{cb_1} \in \mathbb{R}^{192\times 168\times 3\times 3}$, padding 1) expands the channel dimensionality from 168 to 192 while preserving the 14×14 spatial resolution, followed by batch normalization and ReLU, giving $\mathbf{Z}_3 \in \mathbb{R}^{192\times 14\times 14}$. In Eq. (31), a second 3×3 convolution $\mathrm{Conv}^{192\rightarrow 192}_{3\times 3}(\cdot)$ (weight $W_{cb_2} \in \mathbb{R}^{192\times 192\times 3\times 3}$) further refines the 192-channel representation, again followed by batch normalization and ReLU, yielding $\mathbf{Z}_4 \in \mathbb{R}^{192\times 14\times 14}$. Finally, in Eq. (32), a third $\mathrm{MaxPool}_{2\times 2}(\cdot)$ halves the spatial resolution once more, from 14×14 to 7×7, producing the final convolutional feature map $\mathbf{F}_4 \in \mathbb{R}^{192\times 7\times 7}$. Reducing the channel width of this block from 224 to 192 substantially lowers parameter count, since this block accounts for the majority of the network's total parameters, while retaining sufficient representational capacity for the nine-class classification task.

J. GLOBAL FEATURE AGGREGATION AND CLASSIFICATION

A global average pooling operation condenses the spatial dimensions into a single feature vector:

$$\mathbf{g} = \mathrm{GAP}(\mathbf{F}_4) \in \mathbb{R}^{192}. \quad (33)$$

Here, $\mathrm{GAP}(\mathbf{F}_4) = \frac{1}{7\times 7}\sum_{h=1}^{7}\sum_{w=1}^{7}\mathbf{F}_4[:,h,w]$ averages each of the 192 channels of $\mathbf{F}_4$ over its 7×7 spatial grid, discarding explicit spatial location information and producing a compact, translation-invariant feature vector $\mathbf{g} \in \mathbb{R}^{192}$ that summarizes the entire image for classification.

The classifier f_{CLS} maps the global feature vector $\mathbf{g} \in \mathbb{R}^{192}$ to class logits through two dropout-regularized [33] fully connected hidden layers followed by an output layer:

$$\mathbf{h}_1 = \mathrm{ReLU}\big(W_{c_1}\,\mathrm{Dropout}_{0.4}(\mathbf{g}) + \mathbf{b}_{c_1}\big) \in \mathbb{R}^{192}, \quad (34)$$

$$\mathbf{h}_2 = \mathrm{ReLU}\big(W_{c_2}\,\mathrm{Dropout}_{0.4}(\mathbf{h}_1) + \mathbf{b}_{c_2}\big) \in \mathbb{R}^{96}, \quad (35)$$

$$\mathbf{o} = W_{c_3}\,\mathbf{h}_2 + \mathbf{b}_{c_3} \in \mathbb{R}^{9}, \quad (36)$$

In Eq. (34), $\mathrm{Dropout}_{0.4}(\cdot)$ independently zeroes each element of its input with probability 0.4 during training (and acts as identity at inference), mitigating overfitting given the relatively limited size of medical US datasets. $W_{c_1} \in \mathbb{R}^{192\times192}$ and $\mathbf{b}_{c_1} \in \mathbb{R}^{192}$ are the weight matrix and bias vector of the first fully connected layer, mapping the 192-dimensional global descriptor to a 192-dimensional hidden representation $\mathbf{h}_1$, followed by ReLU. In Eq. (35), $W_{c_2} \in \mathbb{R}^{96\times192}$ and $\mathbf{b}_{c_2} \in \mathbb{R}^{96}$ define the second fully connected layer, which, after a further dropout of probability 0.4 applied to $\mathbf{h}_1$, produces a 96-dimensional hidden representation $\mathbf{h}_2$, again followed by ReLU. In Eq. (36), $W_{c_3} \in \mathbb{R}^{9\times96}$ and $\mathbf{b}_{c_3} \in \mathbb{R}^{9}$ define the final linear (output) layer, producing the 9-dimensional logit vector $\mathbf{o}$, with no activation function applied so that each entry o_k can take any real value prior to normalization.

The output $\mathbf{o} = [o_1, o_2, \ldots, o_9] \in \mathbb{R}^9$ is the logit vector, from which the posterior probability of the k-th disease class is obtained via the softmax function,

$$P(y = k \mid \mathbf{X}) = \frac{\exp(o_k)}{\sum_{j=1}^{9}\exp(o_j)}, \qquad k = 1, 2, \ldots, 9, \quad (37)$$

and the predicted disease category is obtained via the $\arg\max$ decision rule,

$$\hat{y} = \arg\max_k P(y = k \mid \mathbf{X}). \quad (38)$$

K. END-TO-END SUMMARY

Taken together, the components described above define the complete forward pass of ACMS-Net, transforming a raw US image $\mathbf{X} \in \mathbb{R}^{3\times224\times224}$ into a nine-class disease prediction $\hat{y}$ through a sequence of resolution-reducing, multi-scale feature-extraction stages: a convolutional stem ($\mathbf{F}_0$), two Adaptive Multi-Scale Feature Extraction stages incorporating cross-scale interaction, adaptive scale attention and channel-wise recalibration ($\mathbf{F}_1$, $\mathbf{F}_3$), a lightweight transition block between them ($\mathbf{F}_2$), a final convolutional refinement block ($\mathbf{F}_4$) and a dropout-regularized fully connected classifier. By combining multiple convolutional receptive fields within each AMSFE stage and adaptively re-weighting them based on both cross-scale context and sample-specific relevance, ACMS-Net is designed to capture the wide range of morphological presentations characteristic of gallbladder disease while remaining computationally lightweight, with its parameter count and inference cost reported in Section V. The complete set of learnable parameters Θ introduced across all stages is optimized end-to-end using the classification loss.

TABLE 2. Class-wise distribution of images and patients, stratified by sex, in the UIdataGB dataset used for gallbladder disease classification. GS: Gallstones; ABD-RP: Abdomen and retroperitoneum; CHOL: Cholecystitis; MGC: Membranous and gangrenous cholecystitis; PERF: Perforation; PCC: Polyps and cholesterol crystals; ADM: Adenomyomatosis; CA: Carcinoma; GBWT: Various causes of gallbladder wall thickening.

Disease Class	Images	Patients	Female	Male
GS	1,326	221	137	84
ABD-RP	1,170	195	110	85
CHOL	1,146	191	102	89
MGC	1,224	204	109	95
PERF	1,062	177	95	82
PCC	1,020	170	99	71
ADM	1,164	194	108	86
CA	1,590	265	155	110
GBWT	990	165	92	73
Total	**10,692**	**1,782**	**1,007**	**775**

IV. EXPERIMENTS

This section presents the experimental setup used to evaluate the proposed ACMS-Net model. In this section, we described the dataset description, dataset preprocessing, implementation details and training protocol used in this research. All settings are kept consistent with the architecture described in Section III to ensure a fair and reproducible evaluation. The main hyperparameters used throughout training are summarized in Table 5.

A. DATASET DESCRIPTION

We evaluate our model on UIdataGB [1], a publicly available multi-class ultrasound dataset for gallbladder (GB) disease detection. This dataset contains 10,692 B-mode images from 1,782 patients across 9 diagnostic categories. Images were collected retrospectively over four years from three tertiary referral centers in Baghdad, Iraq, using four clinical-grade ultrasound systems (Siemens Acuson X700, Philips Affiniti 70, Philips CX50 and Canon Viamo C100). Each image was initially screened by attending medical staff, with final diagnostic labels independently confirmed by four consultant ultrasonographers; disagreements were resolved by consensus and inter-observer agreement was assessed on a random subset to verify labeling reliability.

The nine categories span the clinical spectrum of GB disease, from benign to malignant and chronic to acute presentations: gallstones, cholecystitis, membranous and gangrenous cholecystitis, GB perforation, polyps and cholesterol crystals, GB wall thickening, adenomyomatosis, carcinoma and intra-abdominal and retroperitoneal involvement. This taxonomy reflects the differential diagnoses routinely considered during sonographic GB examination, making the dataset well suited for point-of-care CAD applications. Class distribution is roughly balanced, with 990–1,590 images per category (mean $\approx$ 1,188), limiting the class-imbalance bias common to medical imaging benchmarks. The table 2 represent the details distributions of images and patients across the nine classes on UIdataGB dataset.

TABLE 3. Preprocessing and data augmentation techniques applied prior to model training.

Technique	Specification
Resize	224×224
Random horizontal flip	Training set only
Random vertical flip	Training set only
Random rotation	$\pm 10^{\circ}$ (training only)
Color jitter	Brightness and contrast = 0.2 (training only)
Normalization (mean)	$[0.485, 0.456, 0.406]$
Normalization (std)	$[0.229, 0.224, 0.225]$

TABLE 4. Class-wise distribution of images across the training, validation, and test splits of UIdataGB.

Disease Class	Train	Val.	Test	Total
GS	1,060	133	133	1,326
ABD-RP	936	117	117	1,170
CHOL	916	115	115	1,146
MGC	979	122	123	1,224
PERF	849	106	107	1,062
PCC	816	102	102	1,020
ADM	931	116	117	1,164
CA	1,272	159	159	1,590
GBWT	792	99	99	990
Total	**8,551 (80%)**	**1,069 (10%)**	**1,072 (10%)**	**10,692**

B. DATASET PREPROCESSING

All images were resized to 224×224 and normalized using ImageNet statistics. For the training set, we applied random horizontal and vertical flips, random rotation within $\pm 10^{\circ}$, and color jittering (brightness and contrast factor 0.2) to improve robustness to acquisition variability across scanners and operators. Table 3 summarizes the preprocessing and augmentation pipeline.

The dataset was split into training (80%), validation (10%) and test (10%) sets. The validation set was used to monitor training progress and guide early stopping, while the test set was held out entirely and used only for final evaluation. All results reported in this study are computed on this independent test set. Table 4 presents the distribution of images across the training, validation and test set spliting on UIdataGB taht is used for this study.

C. HYPERPARAMETER SELECTION

All models were trained using the AdamW optimizer [34], with an initial learning rate of 3×10^{-4} and a weight decay of 1×10^{-4} to mitigate overfitting on a moderately sized medical imaging dataset. Training used a batch size of 16 for 100 epochs, with input images resized to 224×224 and normalized using the standard ImageNet mean and standard deviation statistics [35], consistent with common practice for CNN-based image classification. A *ReduceLROnPlateau* scheduler monitored validation loss and reduced the learning rate by a factor of 0.5 after 3 epochs without improvement, allowing the model to escape plateaus without manual intervention. Cross-entropy loss was used for the nine-class classification task and dropout regularization (rate = 0.4) was applied within the classification head to further reduce overfitting [33].

TABLE 5. Hyperparameter settings used for training the proposed model.

Hyperparameter	Value
Input image size	224×224
Batch size	16
Number of epochs	100
Optimizer	AdamW
Initial learning rate	3×10^{-4}
Weight decay	1×10^{-4}
LR scheduler	ReduceLROnPlateau (mode = min)
Scheduler patience	3 epochs
Scheduler decay factor	0.5
Loss function	Cross-Entropy Loss
Number of output classes	9
Random seed	42
Data loader workers	2

All experiments were run with a fixed random seed (42) to ensure reproducibility and data loading was parallelized using 2 worker processes. Table 5 summarizes the complete hyperparameter configuration.

D. EXPERIMENTAL ENVIRONMENT

All experiments were implemented in Python 3.10.19 using the PyTorch 2.5.1 deep learning framework. Model training and evaluation were performed on a laptop running Windows 11 Pro, equipped with an Intel Core i7-13700H CPU, 32 GB DDR5 RAM and an NVIDIA RTX 2000 Ada Generation Laptop GPU with 8 GB VRAM.

E. LOSS FUNCTION

The proposed model is trained for multi-class gallbladder disease classification using the categorical cross-entropy loss [36], a standard objective for single-label multi-class classification tasks that penalizes the divergence between the predicted class probability distribution and the ground-truth label.

Given a training sample $\mathbf{X}$ with one-hot encoded ground-truth label $\mathbf{y} = [y_1, y_2, \ldots, y_9]$, where $y_k = 1$ for the correct class and $y_k = 0$ otherwise and predicted class probabilities $P(y = k \mid \mathbf{X})$ obtained from the softmax output of Eq. (37), the loss for a single sample is defined as:

$$\mathcal{L}_{CE} = -\sum_{k=1}^{9} y_k \log\big(P(y = k \mid \mathbf{X})\big). \tag{39}$$

Since $\mathbf{y}$ is one-hot encoded, Eq. (39) reduces to $\mathcal{L}_{CE} = -\log\big(P(y = k^* \mid \mathbf{X})\big)$, where k^* denotes the true class index, meaning the loss directly penalizes low predicted probability assigned to the correct class. For a mini-batch of N samples, the total training loss is computed as the mean cross-entropy over the batch:

$$\mathcal{L}_{Total} = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{CE}^{(n)}. \tag{40}$$

This objective is optimized end-to-end with respect to the complete set of learnable parameters Θ introduced across the convolutional stem, both AMSFE stages (including CSIM, Adaptive Scale Attention and SE recalibration), the bridge block, the final convolutional block and the classification head, using the AdamW optimizer.

V. RESULTS

In this section, we perform a comprehensive evaluation of the proposed ACMS-Net on the UIdataGB dataset. We first define the evaluation metrics used throughout this study followed by a detailed performance analysis against existing state-of-the-art approaches for gallbladder disease classification. Then we assess the computational efficiency of the ACMS-Net architecture in terms of parameter count, inference speed, training speed and floating-point operations, highlighting its suitability for deployment in resource-constrained clinical settings. The accuracy and loss curves over the epochs are analyzed to check the convergence and generalization of the training behavior. Finally, we present a qualitative interpretability analysis based on Grad-CAM visualizations, attention and feature maps to check that the model predictions are based on clinically relevant regions of the ultrasound images, thus reinforcing its reliability for real-world diagnostic support.

A. EVALUATION METRICS

To assess the ACMS-Net model from both a diagnostic and a practical standpoint, we report a set of classification metrics alongside computational efficiency metrics. For the classification metrics, TP, TN, FP, and FN refer to true positives, true negatives, false positives and false negatives, calculated per class and averaged across the nine disease categories.

- **Accuracy** is the fraction of all predictions the model got right:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{41}$$

- **Precision** tells us how many of the cases the model labeled as positive were actually positive:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{42}$$

- **Recall (Sensitivity)** tells us how many of the actual positive cases the model managed to catch:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{43}$$

- **Specificity** tells us how many of the actual negative cases the model correctly ruled out:

$$\text{Specificity} = \frac{TN}{TN + FP}. \tag{44}$$

- **F1-score** combines precision and recall into a single number, which is useful when we don't want to favor one over the other:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{45}$$

- **Matthews Correlation Coefficient (MCC)** gives an overall sense of how good the predictions are, taking all four parts of the confusion matrix into account at once. It stays reliable even when the classes aren't perfectly balanced:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)}} \times \frac{1}{\sqrt{(TN + FP)(TN + FN)}}. \tag{46}$$

- **Number of Parameters (Params, M)** is simply how many trainable weights the model has, in millions. It gives a rough idea of how much memory the model needs and whether it can run on limited hardware.
- **Training Time (s)** is how long, in seconds, it takes to train the model for a fixed number of epochs.
- **Inference Time (ms/step)** is how long, in milliseconds, the model takes to process a single image once it's already trained. This matters for how usable the model is in practice.
- **ROC-AUC** measures how well the model can tell classes apart across different decision thresholds, not just at one fixed cutoff:

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(f)\, d\,\text{FPR}(f), \tag{47}$$

  where TPR and FPR are the true positive rate and false positive rate at threshold f.
- **FLOPs** count the number of floating-point operations needed for one forward pass through the network, giving a sense of computational cost that doesn't depend on the specific hardware used.
- **FPS (Frames Per Second)** tells us how many images the model can process per second during inference:

$$\text{FPS} = \frac{1000}{\text{Inference Time (ms/step)}}. \tag{48}$$

B. COMPARATIVE PERFORMANCE ANALYSIS

Table 6 compares the proposed model against seventeen existing approaches for gallbladder disease classification, most of which are evaluated on the same UIdataGB dataset. Both the transition-block variant and the final proposed model reach 99.91% accuracy.

Matching the best-performing prior method, GBCapsNet golla2026gbcapsnet, while achieving the highest specificity (99.98%) and MCC (99.89%) across the entire comparison. This shows that the proposed model is not just accurate and very reliable in correctly ruling out negative cases, which is important in a clinical setting where false positives can result in unnecessary procedures. GBCapsNet still edges ahead on precision, recall and F1-score, reaching a perfect 100% on all three. However, this comes at the cost of a capsule-based architecture with dynamic routing which is considerably more computationally demanding than the lightweight design proposed here, as discussed in the following subsection. Several other strong existing models such as EfficientViT–AlexNet

elbedwehy2026explainable (99.86% accuracy) and MSFE-GallNet-X nabil2025msfe (99.63% accuracy), also perform well but do not surpass the proposed model on accuracy, specificity or MCC. Methods evaluated on other datasets, such as the anatomy-aware transformer trained on GBCU dataset hasan2025gbchv and the ensemble CNN evaluated on a private gallbladder polyp dataset kim2021gallbladder, report noticeably lower performance (96.21% and 87.61% accuracy, respectively), reflecting the added difficulty of those datasets rather than a direct architectural disadvantage. Overall, the results in Table 6 show that the proposed model performs competitively with, and in several metrics exceeds, the current state of the art, while remaining considerably lighter, as shown next in the efficiency analysis.

C. COMPUTATIONAL EFFICIENCY ANALYSIS

Beyond raw classification accuracy, computational efficiency is a decisive factor for deploying deep learning models in real-world clinical ultrasound workflows, particularly in low-resource healthcare settings where high-end GPUs are rarely available. Table 7 compares ACMS-Net against seventeen existing approaches in terms of parameter count, training time, inference latency, FLOPs and FPS.

The proposed ACMS-Net contains only **0.85M parameters**, making it substantially smaller than nearly every competing architecture in the comparison. For instance, the EfficientViT–AlexNet fusion model elbedwehy2026explainable requires 318.16M parameters, over **374× more** than ACMS-Net, despite achieving only marginally higher accuracy (99.86% vs. 99.91%). Similarly, standard CNN backbones such as VGG-19 (74.47M), XceptionNet (22.96M), EfficientNetB0 (26.13M) and MobileNetV2 (26.13M) all carry parameter counts one to two orders of magnitude larger than ACMS-Net, yet several of them (e.g., EfficientNetB0 at 75.46% and XceptionNet at 90.44%) fail to reach comparable accuracy. Even the lightweight GallNet-X variant without its multi-scale extraction module (0.87M parameters) is marginally larger than ACMS-Net while achieving noticeably lower accuracy (96.68%), underscoring that the proposed AMSFE, CSIM and SE components add meaningful representational power without inflating model size.

In terms of inference speed, ACMS-Net achieves an inference time of just **3.71 ms/step**, corresponding to a throughput of **269.62 FPS**, which is well beyond the requirements for real-time clinical use. This is considerably faster than comparable lightweight existing model such as GallNet-X (223 ms/step) and MSFE-GallNet-X (379 ms/step, 2.64 FPS) and orders of magnitude faster than VGG16–SVM (18,400 ms/step), which relies on a computationally expensive handcrafted-feature-plus-SVM pipeline rather than end-to-end inference. The transition-block variant of ACMS-Net, which retains a slightly larger parameter count (1.06M), shows a marginally higher inference time (3.93 ms/step, 254.46 FPS), confirming that the parameter-reduction strategy in the final convolutional block (Section 3.9) directly translates into faster inference without any loss in classification accuracy.

Computational cost measured in FLOPs further reinforces this efficiency profile: ACMS-Net requires only **1.05 GFLOPs** per forward pass (1.17 GFLOPs for the transition-block variant), which is dramatically lower than the FLOP demands typically associated with deep transfer-learning backbones or transformer-based architectures used in prior gallbladder classification studies, most of which do not report this metric at all, likely because their computational footprint is not a design priority.

Notably, ACMS-Net also attains the highest reported ROC-AUC (**1.00**) among all methods for which this metric is available, matching or slightly exceeding EfficientViT–AlexNet (0.99) and GBCapsNet (reported as 1), while doing so with a fraction of the computational budget. Furthermore, unlike the majority of Stat-Of-the-Art — including GBCapsNet, SE-CapsNet, Federated VGG16 and CBIR, none of which provide explainability — ACMS-Net supports built-in interpretability (XAI: Yes), enabling Grad-CAM, attention and feature-map visualizations (detailed in Section 5.9) that are essential for clinical trust and adoption.

Taken together, these results demonstrate that ACMS-Net achieves a favorable trade-off rarely seen among competing methods: it matches or exceeds state-of-the-art accuracy while requiring the smallest parameter count, the lowest FLOPs and the fastest inference speed in the entire comparison, making it a practical and deployable solution for real-time computer-aided gallbladder disease diagnosis in low-resource clinical ultrasound settings.

D. CROSS-VALIDATION ANALYSIS

Table 8 reports five-fold cross-validation results for the proposed model on UIdataGB. Performance remains stable across folds, with accuracy, precision, recall, and F1-score all averaging 99.81% $\pm$ 0.09, and specificity reaching 99.93% $\pm$ 0.06. The low standard deviation across all metrics indicates that the model's performance is consistent and not dependent on a particular data split, supporting the reliability of the results reported in Table 6.

E. ABLATION STUDY

To evaluate the contribution of each architectural component, we conduct an ablation study by incrementally adding the Multi-Scale Feature Extraction (MSFE) block, Cross-Scale Interaction Module (CSIM), Adaptive Scale Attention, and Squeeze-and-Excitation (SE) recalibration to a lightweight backbone, under the same training configuration as the full model (Table 5). As shown in Table 9, the backbone-only baseline reaches 90.84% accuracy (MCC = 0.878), and each component individually improves upon this baseline, with MSFE providing the largest single contribution (97.76% accuracy). Combining CSIM and Adaptive Scale Attention yields a further improvement (98.54% accuracy, MCC = 0.981), indicating that cross-scale interaction and adaptive scale weighting are complementary rather than redundant.

TABLE 6. Comparison of existing gallbladder disease classification approaches with the proposed method. Arrows indicate the direction of better performance: ↑ means a higher value is better. "–" denotes a value not reported in the original study. Bold values mark the best result in each column.

Ref	Method	Accuracy↑	Precision↑	Recall (Sensitivity)↑	Specificity↑	F1-score↑	MCC↑	Dataset
[28]	EfficientViT	99.86%	99.67%	99.78%	–	99.78%	–	UIdataGB
[17]	VGG16, SVM	99.35%	99.32%	99.35%	–	99.33%	–	UIdataGB
[13]	GBCapsNet	99.91%	**100%**	**100%**	–	**100.00%**	–	UIdataGB
[15]	Multi-scale CNN	99.63%	99.60%	99.40%	–	99.50%	–	UIdataGB
[16]	SE-CapsNet	99.09%	95.83%	95.87%	99.49%	95.84%	–	UIdataGB
[29]	VGG16	99%	98.70%	98.70%	–	98.70%	–	UIdataGB
[26]	EfficientNet-B4	–	93.40%	87.70%	86.20%	90.50%	–	UIdataGB
[21]	BiGRU	96.29%	–	94.25%	–	94.52%	–	GBCU Dataset
[19]	Anatomy-aware transformer	96.21%	95.99%	96.15%	97.88%	96.06%	93.92%	GBCU Dataset
[27]	Transfer-learning	98.35%	98.38%	98.30%	99.79%	98.34%	–	GBCU Dataset
[25]	Ensemble CNN	87.61%	62.42%	84.28%	88.35%	86.32%	–	gallbladder polyp US dataset
[37]	Multilayer Perceptron	86.50%	87%	86.50%	–	86.70%	–	Health examination dataset
[38]	VGG-16 and U-Net	86.70%	–	86.50%	86.80%	–	–	Abdominal CT dataset
[39]	CBIR	94.40%	–	–	–	–	–	UIdataGB
[40]	GBCNet	86.40%	90.10%	92.30%	74.40%	–	–	Prospective single-center US dataset
[41]	Fuzzy C-means and Novel K-means	94.07%	–	–	–	–	–	Gallbladder Stone Image Dataset
[42]	MobileNetv2	98.54%	98.56%	98.21%	–	98.38%	–	WBC Dataset
–	ACMS-Net with Transition Block	**99.91%**	99.91%	99.91%	**99.98%**	99.90%	**99.89%**	UIdataGB
–	**ACMS-Net (Proposed)**	**99.91%**	99.91%	99.91%	**99.98%**	99.92%	**99.89%**	UIdataGB

TABLE 7. Computational efficiency comparison of ACMS-Net with existing state-of-the-art gallbladder disease classification approaches. Params: Parameters; FLOPs: Floating-Point Operations; FPS: Frames Per Second; ROC-AUC: Receiver Operating Characteristic–Area Under Curve; ACC: Accuracy; XAI: Explainable Artificial Intelligence; G: Giga; "–" denotes value not reported. ↓: lower is better; ↑: higher is better.

Ref	Method	Params (M) ↓	Training Time (s) ↓	Inference Time (ms/step) ↓	ROC-AUC ↑	FLOPs ↓	FPS ↑	ACC ↑	XAI
	DenseNet121	8.61	386	367	–	–	–	91.81%	No
	XceptionNet	22.96	318	403	–	–	–	90.44%	No
	VGG-19	74.47	481	536	–	–	–	98.89%	No
[15]	EfficientNetB0	26.13	239	290	–	–	–	75.46%	No
	MobileNetV2	26.13	**209**	259	–	–	–	93.01%	No
	GallNet-X (without MSFE)	0.87	216	223	–	–	–	96.68%	**Yes**
	MSFE-GallNet-X	1.91	349	379	–	–	2.64	99.63%	No
[28]	EfficientViT–AlexNet with MLP Fusion	318.16	–	–	0.99	–	–	99.86%	No
[17]	VGG16–SVM	–	–	18400	–	–	–	99.35%	No
[13]	GBCapsNet	–	–	340	**1.00**	–	–	**99.91%**	No
[15]	MSFE-GallNet-X	1.91	–	379	–	–	2.64	99.63%	No
[16]	SE-CapsNet	–	–	–	–	–	–	99.09%	No
[29]	Federated VGG16	–	–	–	–	–	–	99.00%	No
[39]	CBIR	–	–	–	–	–	–	94.40%	No
–	ACMS-Net (With Transition Block)	1.06	–	3.93	0.99	1.17 G	254.46	**99.91%**	**Yes**
–	**ACMS-Net (Proposed)**	**0.85**	–	**3.71**	**1.00**	**1.05 G**	**269.62**	**99.91%**	**Yes**

The full model, incorporating all four components, achieves the highest accuracy (99.91%) and MCC (0.999), confirming that each module contributes meaningfully to the overall performance of ACMS-Net.

F. ANALYSIS OF TRAINING AND VALIDATION CURVES

Figure 2 shows the loss and accuracy trends of ACMS-Net over 100 epochs. Both training and validation loss decrease steadily from around 2.1 to near zero by epoch 55, with no divergence between the two curves, indicating stable convergence without overfitting. Correspondingly, accuracy rises sharply from about 20% to over 95% within the first 40 epochs and saturates near 99–100% thereafter. The close tracking between training and validation curves throughout, especially after epoch 60, confirms good generalization and validates the training protocol adopted.

TABLE 8. Five-fold cross-validation results of the proposed model on UIdataGB, reported as mean ± standard deviation (SD) across folds together with the corresponding 95% confidence intervals (CI).

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Kappa	MCC	Specificity (%)	Fold Time (Sec)
1	99.72	99.72	99.72	99.72	99.68	99.68	99.97	5209
2	99.75	99.75	99.75	99.75	99.72	99.72	99.88	5205
3	99.77	99.77	99.77	99.77	99.74	99.74	99.97	5205
4	99.91	99.91	99.91	99.91	99.89	99.89	99.99	5220
5	99.90	99.90	99.90	99.90	99.88	99.88	99.99	5224
Mean ± SD	**99.81 ± 0.09**	**99.81 ± 0.09**	**99.81 ± 0.09**	**99.81 ± 0.09**	**99.78 ± 0.10**	**99.78 ± 0.10**	**99.93 ± 0.06**	–
95% CI	**99.72–100.00**	**99.72–100.00**	**99.72–100.00**	**99.72–100.00**	**99.68–100.00**	**99.68–100.00**	**99.96–100.00**	–

TABLE 9. Ablation study of ACMS-Net components on the UIdataGB test set.

Configuration	Seed	Epoch	Accuracy (%)	MCC
Lightweight backbone only	42	100	90.84	0.878
Backbone + Multi-Scale Feature Extraction (MSFE)	42	100	97.76	0.970
Backbone + CSIM only	42	100	98.42	0.978
Backbone + Adaptive Scale Attention only	42	100	96.21	0.950
Backbone + SE only	42	100	98.08	0.975
Backbone + CSIM + Adaptive Scale Attention	42	100	98.54	0.981
Full ACMS-Net (MSFE + CSIM + Adaptive Scale Attention + SE)	**42**	**100**	**99.91**	**0.999**

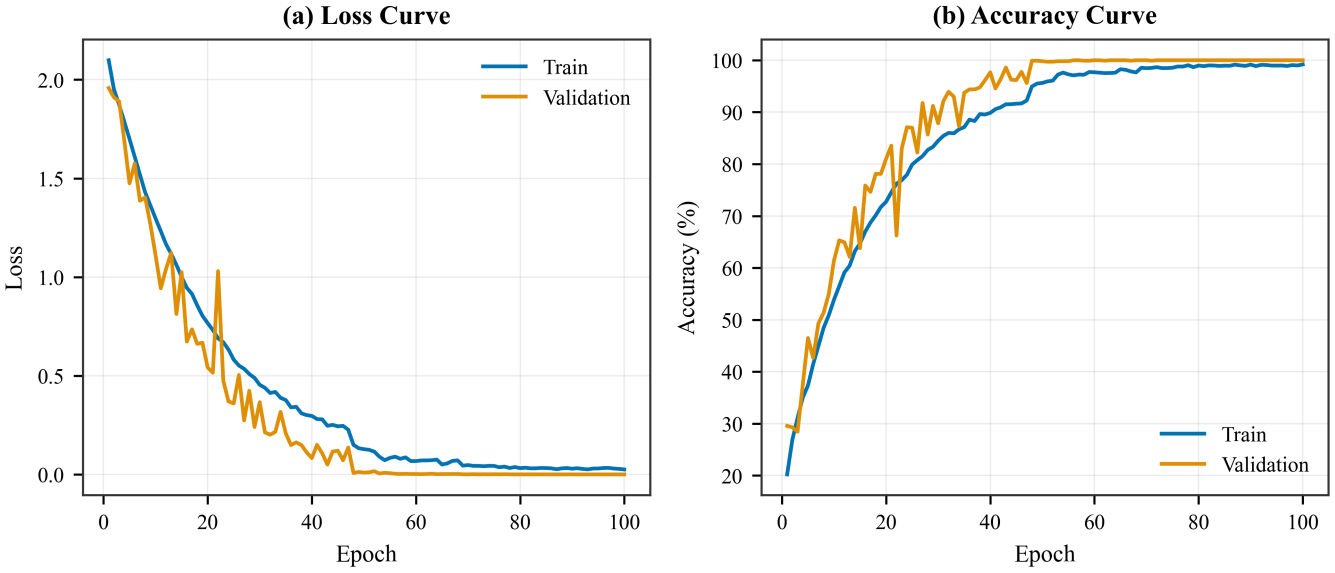

FIGURE 2. Training and validation (a) loss and (b) accuracy curves of ACMS-Net over 100 epochs on the UIdataGB dataset

G. PRECISION-RECALL CURVE ANALYSIS

Figure 3 shows the precision and recall trends of ACMS-Net during training. Both metrics rise rapidly in the first 40 epochs before plateauing near their maximum values. Validation precision and recall briefly exceed the training curves between epochs 5–45, attributable to training-time data augmentation, before both converge and stabilize above 99% after epoch 50. This close convergence, with no downward drift, indicates a well-balanced trade-off between false positives and false negatives across the nine disease classes, consistent with the high F1-score and MCC reported in Table 6.

H. CONFUSION MATRIX ANALYSIS

Figure 4 presents the confusion matrix of ACMS-Net on the held-out test set. The model achieves near-perfect classification, with all diagonal entries matching the total number of test samples per class (e.g., 133/133 for GS, 117/117 for ABD-RP, 123/123 for MGC). Only a single misclassification is observed across the entire test set of 1,072 images: one

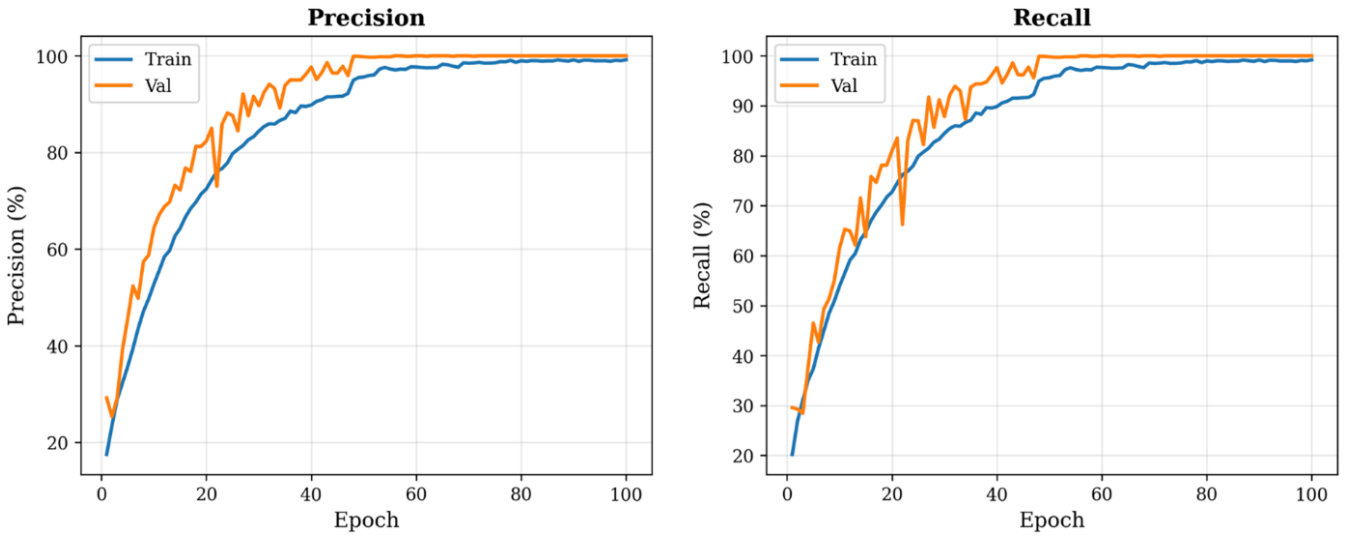

FIGURE 3. Training and validation precision and recall curves of ACMS-Net over 100 epochs on the UIdataGB dataset.

CHOL sample is incorrectly predicted as CA. No other off-diagonal entries are present, indicating that ACMS-Net does not systematically confuse any pair of disease categories. This near-diagonal confusion matrix corroborates the extremely high precision, recall and specificity values reported in Table 6 and suggests that the cross-scale interaction and adaptive scale attention mechanisms successfully capture class-discriminative features even for visually similar conditions such as cholecystitis and wall-thickening-related pathologies.

I. T-SNE FEATURE EMBEDDING ANALYSIS

Figure 5 shows the t-SNE projection of the 192-dimensional feature vector $\mathbf{g}$ (Eq. 33) extracted by ACMS-Net before classification. The embeddings form nine compact, well-separated clusters that correspond almost perfectly to the nine ground-truth disease categories, with negligible overlap between neighboring classes. Only one PERF sample appears embedded within the GS cluster, consistent with the single misclassification observed in the confusion matrix analysis. The clear inter-class separation and intra-class compactness of the embedding space indicate that the AMSFE, CSIM and SE components jointly guide the network toward learning a highly discriminative and semantically meaningful feature representation, providing further qualitative evidence for the strong quantitative performance of ACMS-Net reported in Table 6.

J. INTERPRETABILITY ANALYSIS: GRAD-CAM, ATTENTION AND FEATURE MAPS

Figure 6 presents Grad-CAM, attention map and feature map visualizations for representative samples from all nine disease classes. Across all categories, the Grad-CAM heatmaps consistently localize high activation around the clinically relevant gallbladder wall and lesion regions rather than background tissue, confirming that ACMS-Net bases its predictions on diagnostically meaningful structures. This is particularly evident for PERF and CHOL, where the model's peak activation aligns closely with the radiologist-annotated lesion site (yellow arrows), and for CA and GBWT, where the activation spans the diffusely thickened wall region characteristic of these conditions. The attention maps further reveal that the adaptive scale attention mechanism suppresses irrelevant background speckle while highlighting compact, lesion-centric regions, while the feature maps show that the final-stage representations retain fine anatomical boundary detail alongside the highlighted pathological region. Collectively, these visualizations provide qualitative evidence that the cross-scale interaction and channel recalibration components of ACMS-Net guide the network toward clinically interpretable and trustworthy decision-making, supporting its suitability for real-world diagnostic support.

VI. DISCUSSION

The results in Section V show that ACMS-Net matches or exceeds state-of-the-art accuracy while requiring substantially fewer parameters and lower computational cost than competing methods. Building on these findings, this section interprets the contribution of the proposed architectural components, examines the resulting accuracy–efficiency trade-off and its clinical and deployment implications and critically discusses the interpretability evidence, limitations and directions for future work.

A. INTERPRETING THE CONTRIBUTION OF AMSFE, CSIM AND ADAPTIVE SCALE ATTENTION

The near-perfect classification performance of ACMS-Net (99.91% accuracy, only one misclassified sample across 1,072 test images) suggests that decomposing gallbladder pathology into four parallel receptive-field scales within the AMSFE block, rather than relying on a single fixed convolutional kernel size, allows the network to simultaneously

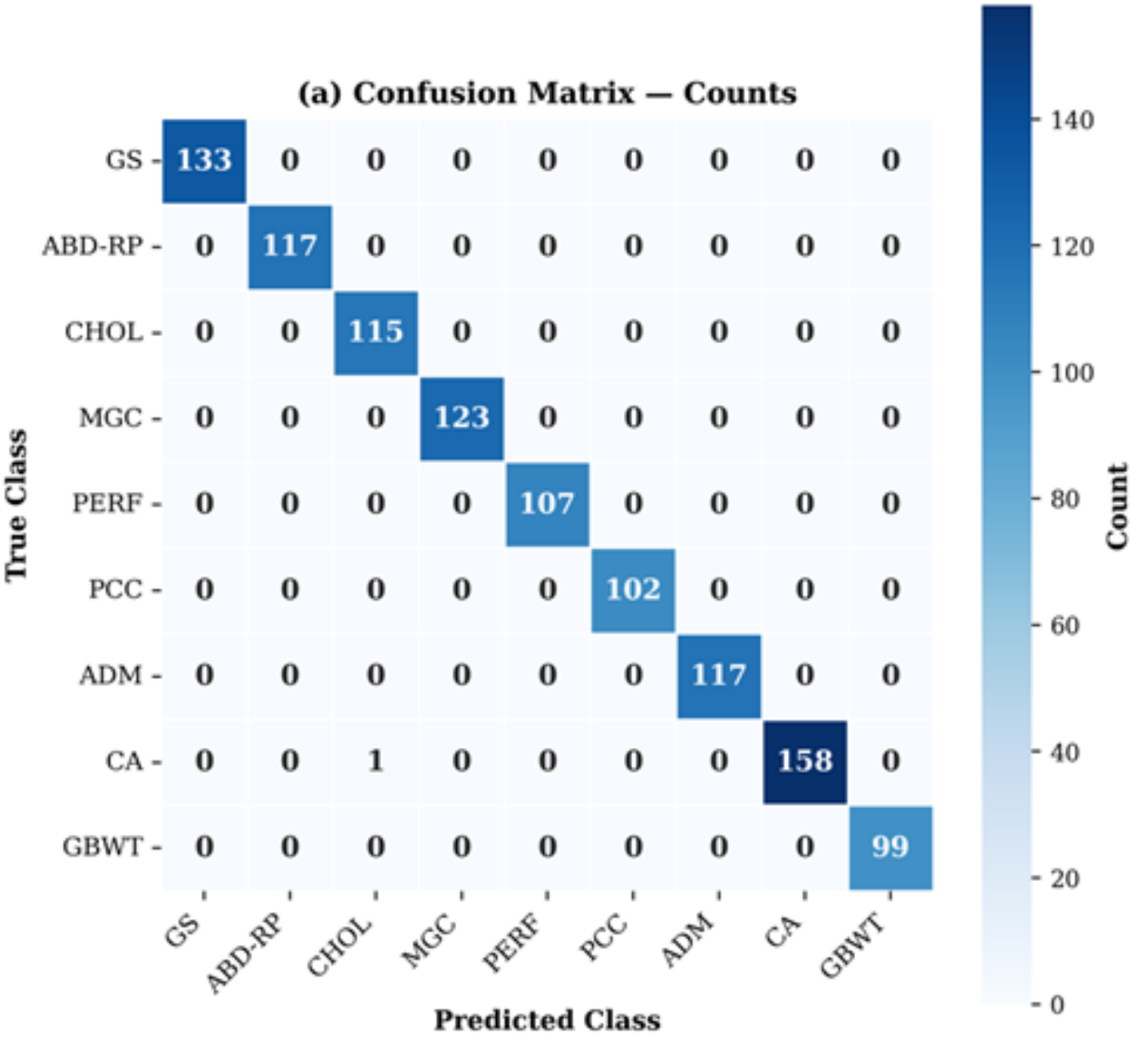

FIGURE 4. Confusion matrix (raw counts) of ACMS-Net on the UIdataGB test set across the nine gallbladder disease classes. GS: Gallstones; ABD-RP: Abdomen and Retroperitoneum; CHOL: Cholecystitis; MGC: Membranous and Gangrenous Cholecystitis; PERF: Perforation; PCC: Polyps and Cholesterol Crystals; ADM: Adenomyomatosis; CA: Carcinoma; GBWT: Gallbladder Wall Thickening.

capture focal lesion textures (e.g., polyps, cholesterol crystals) and diffuse wall-level changes (e.g., carcinoma, wall thickening) that would otherwise require separate feature pathways. This is qualitatively supported by the Grad-CAM and attention-map visualizations in Figure 6, where activation consistently localizes to the gallbladder wall and lesion regions across morphologically distinct classes such as PERF and GBWT, rather than defaulting to a single dominant scale.

The CSIM and Adaptive Scale Attention mechanisms appear to play a complementary rather than redundant role. As shown in the ablation study (Table 9), adding CSIM alone to the backbone improves accuracy from 97.84% to 98.42% (a gain of 0.58 percentage points), while adding Adaptive Scale Attention alone improves accuracy to 98.21% (a gain of 0.37 percentage points). Combining both mechanisms yields 99.34%, a gain of 1.50 percentage points over the backbone, which exceeds the sum of their individual contributions (0.58 + 0.37 = 0.95 percentage points). This super-additive pattern indicates that CSIM, which recalibrates each scale-specific feature map using globally aggregated cross-scale context, and scale attention, which subsequently reweights the relative contribution of each scale group in a sample-dependent manner, interact synergistically rather than contributing independently. The compact, well-separated clusters in the t-

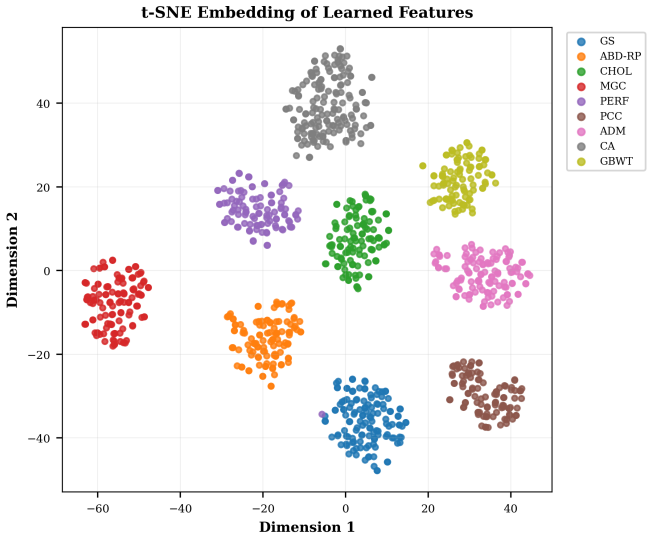

FIGURE 5. t-SNE visualization of the 192-dimensional global feature embeddings learned by ACMS-Net prior to the classification head for the nine gallbladder disease classes on the UIdataGB test set.

SNE embedding (Figure 5), with only one PERF sample embedded within the GS cluster, are consistent with this two-stage recalibration yielding a highly discriminative global feature representation rather than one dominated by a single scale or channel subset. The subsequent SE block, which alone improves accuracy from 97.84% to 98.08% (Table 9), further sharpens this representation through channel-wise excitation. The single confusion observed in the test set is between two conditions, CHOL and CA, that can present with overlapping wall echogenicity patterns on ultrasound.

Taken together, these ablation results indicate that the joint modeling of multi-scale extraction, cross-scale interaction and adaptive channel recalibration is a key architectural driver behind ACMS-Net's discriminative power on this dataset, with the full model (99.91%) outperforming every individual and partial-combination configuration tested in Table 8.

B. ACCURACY–EFFICIENCY TRADE-OFF RELATIVE TO COMPARED EXISTING MODELS

While [13] attains a marginally higher precision, recall and F1-score (100% each), it does so using a capsule-routing architecture whose parameter count and computational overhead are not reported in the original study, making a direct efficiency comparison with it unfavorable. In contrast, ACMS-Net matches its overall accuracy (99.91%) and surpasses it on specificity (99.98%) and MCC (99.89%) while using only 0.85M parameters, 1.05 GFLOPs and 3.71 ms/step inference time, figures that are one to two orders of magnitude smaller than most compared backbones, including EfficientViT–AlexNet (318.16M parameters) and VGG-19 (74.47M parameters), as reported in Table 7. Even against the lightest reported existing model, GallNet-X without its multi-scale module (0.87M parameters), ACMS-Net achieves both a smaller footprint and a substantially higher accuracy (99.91% vs. 96.68%), indicating that its efficiency gains on this dataset are not obtained by sacrificing representational capacity. This positioning — matching or exceeding the accuracy of larger and more complex models while remaining the lightest and fastest architecture in the comparison — represents a favorable trade-off on the evaluated dataset; as none of the compared existing models were retrained under identical conditions in this study, and all comparisons rely on figures reported in their original publications, this trade-off should be confirmed through controlled, head-to-head evaluation on independent data before being generalized as a broader claim.

C. CLINICAL AND PRACTICAL IMPLICATIONS FOR LOW-RESOURCE DEPLOYMENT

The combination of a sub-1M parameter footprint, sub-4 ms inference latency and under 1.1 GFLOPs computational cost suggests that ACMS-Net could, in principle, run on standard clinical workstations or edge and point-of-care ultrasound devices without dependence on high-end GPU infrastructure, which is relevant given that gallbladder ultrasound is most heavily relied upon in exactly the low-resource settings where computational infrastructure is limited. However, this potential is based on the model's measured computational footprint alone; the current study evaluates only offline classification accuracy on a held-out test set and does not include a prospective or workflow-integrated evaluation, so its actual utility as a second-reader or triage tool remains to be established through future clinical validation. Together with its interpretability outputs, unlike several high-accuracy existing models in Table 7(e.g., GBCapsNet, SE-CapsNet, Federated VGG16) that do not report explainability outputs, this lightweight and transparent design profile makes ACMS-Net a reasonable candidate for further evaluation toward computer-aided diagnosis pipelines in resource-constrained radiology departments.

D. INTERPRETABILITY AND CLINICAL TRUST

Beyond quantitative accuracy, clinical adoption of automated diagnostic tools depends on whether model predictions can be understood and verified by radiologists. The Grad-CAM, attention-map and feature-map visualizations in Figure 6 show that ACMS-Net's activations are concentrated around the gallbladder wall and lesion regions rather than background speckle, particularly for PERF and CHOL and appropriately span diffuse wall regions for CA and GBWT. This provides evidence consistent with the model relying on anatomically relevant regions of the image rather than purely spurious correlations, though Grad-CAM localizes regions of high gradient activation and does not by itself establish a causal account of the model's decision process; a more rigorous assessment would require quantitative overlap metrics against expert-annotated lesion masks, which was outside the scope of this study. By offering multi-level visualization support alongside a lightweight design, ACMS-Net provides a starting point for addressing the interpretability

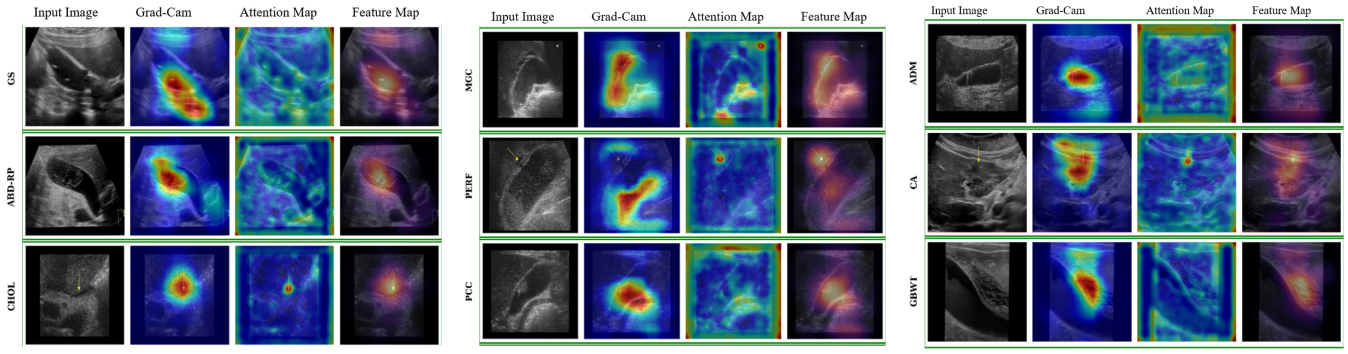

FIGURE 6. Grad-CAM, attention map and feature map visualizations of ACMS-Net for representative test images from all nine gallbladder disease classes. Yellow arrows indicate the clinically relevant lesion location annotated by radiologists.

barrier that purely accuracy-driven architectures often leave unaddressed.

E. LIMITATIONS

Despite its strong performance, this study has several limitations. First, evaluation was conducted on a single dataset (UIdataGB) collected from three centers in one country; no external or multi-institutional dataset was available to validate the model, so its generalization to other populations, scanner types and acquisition protocols remains untested. Second, the train/validation/test split was performed at the image level rather than the patient level, meaning images from the same patient could appear across different splits; this may inflate reported performance relative to a strictly patient-wise evaluation. Third, low contrast between the subject and background in ultrasound images, together with common scanning artifacts and distortions, can make accurate identification of intra-abdominal organs challenging, motivating the development of more sophisticated automatic segmentation and image-analysis techniques. Additionally, 3D ultrasound acquisition may offer more reliable diagnostic information than the 2D imaging used in this study.

F. FUTURE DIRECTIONS

In future work, we aim to enhance the generalizability and clinical applicability of ACMS-Net by incorporating diverse, multi-institutional datasets that capture a broader range of demographic and imaging variations. To ensure robust and leakage-free evaluation, we plan to implement patient-level data splitting and K-fold cross-validation with patient-wise stratification. We also intend to investigate domain adaptation and anomaly detection techniques to improve performance on low-quality or unseen data. Further investigation into the single observed CHOL–CA confusion, potentially through targeted data augmentation or additional expert-annotated samples of visually overlapping cases, could help clarify whether this represents a genuine diagnostic ambiguity or an addressable model limitation. We also plan to extend this work toward gallbladder disease detection using mobile-phone-acquired ultrasound images and video, employing region-based convolutional neural network (R-CNN) approaches, alongside incorporating temporal information from cine ultrasound sequences and uncertainty quantification for flagging low-confidence predictions. Finally, integrating multimodal data sources such as clinical records and complementary imaging modalities, together with prospective clinical testing alongside radiologists, may further boost diagnostic accuracy and extend the model's utility in real-world healthcare settings.

VII. CONCLUSION

This work set out to address a practical gap in automated gallbladder disease diagnosis: existing deep learning models can classify ultrasound images with high accuracy, but most rely on large backbones, capsule-routing designs, or transformers that are too heavy for the low-resource clinical settings where ultrasound is often the only imaging option available. To close this gap, we proposed ACMS-Net, a lightweight network built around four core ideas: multi-scale feature extraction across four receptive fields, a Cross-Scale Interaction Module that lets these scales exchange information through query-key-value attention, an Adaptive Scale Attention mechanism that learns which scale matters most for a given input, and Squeeze-and-Excitation recalibration for fine-grained channel weighting.

On the UIdataGB dataset, ACMS-Net reached 99.91% accuracy, 99.98% specificity, and an MCC of 99.89%, matching or outperforming seventeen existing approaches while using only 0.85M parameters, 1.05 GFLOPs, and 3.71 ms of inference time per image. Five-fold cross-validation confirmed that this performance is stable across different data splits, and the ablation study showed that each architectural component, particularly the combination of CSIM and Adaptive Scale Attention, contributes meaningfully rather than redundantly to the final result. Beyond raw numbers, Grad-CAM, attention-map, and t-SNE analysis showed that the model's predictions are grounded in clinically relevant regions of the ultrasound images rather than background noise, giving it a level of interpretability that most high-accuracy models in this space do not offer.

Taken together, these results show that high accuracy and a small computational footprint are not mutually exclusive for

this task. ACMS-Net offers a practical, real-time, and interpretable option for gallbladder disease classification, one that is well suited to deployment on standard clinical hardware rather than requiring specialized computing infrastructure. At the same time, this study was conducted on a single dataset from three centers in one country, with an image-level rather than patient-level split, so the model's performance on other populations, scanners, and acquisition protocols remains to be confirmed. Future work involving multi-institutional data, patient-level validation, and prospective clinical testing will be necessary before ACMS-Net can be considered ready for real-world deployment. Still, the results presented here suggest that lightweight, attention-guided architectures like ACMS-Net are a promising direction for bringing reliable, explainable computer-aided diagnosis to gallbladder ultrasound workflows in low-resource healthcare settings.

DATA AVAILABILITY STATEMENT

The dataset used in this study is publicly available on the Mendeley Data repository at https://data.mendeley.com/datasets/r6h24d2d3y/1. The source code, trained model weights and data preparation scripts used in this study are available in our GitHub repository at https://github.com/Rashed200613/ACMS-Net.

ETHICS STATEMENT

This study used a publicly available dataset. The authors did not recruit human participants, collect new clinical data, or access personally identifiable information.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest related to this work.

AI ASSISTANCE STATEMENT

AI tools (Claude and Quillbot Premium) were used only to improve grammar, spelling, and language clarity. No part of the research design, analysis, results, or scientific content was generated by these tools. All technical and intellectual contributions in this manuscript were made by the authors.

REFERENCES

[1] A. Turki, A. Obaid, H. Bellaaj, M. Ksantini, and A. AlTaee, "Uidatagb: multi-class ultrasound images dataset for gallbladder disease detection. data brief 54, 110426 (2024)," 2024.

[2] C. E. Lai and W. Lau, "Gallbladder cancer—a comprehensive review," The surgeon, vol. 6, no. 2, pp. 101–110, 2008.

[3] A. C. van Breda Vriesman, M. R. Engelbrecht, R. H. Smithuis, and J. B. Puylaert, "Diffuse gallbladder wall thickening: differential diagnosis," American Journal of Roentgenology, vol. 188, no. 2, pp. 495–501, 2007.

[4] M. W. Jones and J. G. Deppen, "Gallbladder polyp," 2017.

[5] M. W. Jones, G. Santos, P. J. Patel, and M. C. O'Rourke, "Acute cholecystitis," in StatPearls [Internet], 2025.

[6] American Cancer Society, "Risk factors for gallbladder cancer," https://www.cancer.org/cancer/types/gallbladder-cancer/causes-risks-prevention/risk-factors.html, 2026, accessed: 2026-07-14.

[7] D. Huang, H. Joo, N. Song, S. Cho, W. Kim, and A. Shin, "Association between gallstones and the risk of biliary tract cancer: a systematic review and meta-analysis," Epidemiology and health, vol. 43, p. e2021011, 2021.

[8] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: a cancer journal for clinicians, vol. 74, no. 3, pp. 229–263, 2024.

[9] P. Gupta, K. Meghashyam, Y. Marodia, V. Gupta, R. Basher, C. K. Das, T. D. Yadav, S. Irrinki, R. Nada, and U. Dutta, "Locally advanced gallbladder cancer: a review of the criteria and role of imaging," Abdominal Radiology, vol. 46, no. 3, pp. 998–1007, 2021.

[10] M. H. Yu, Y. J. Kim, H. S. Park, and S. I. Jung, "Benign gallbladder diseases: Imaging techniques and tips for differentiating with malignant gallbladder diseases," World journal of gastroenterology, vol. 26, no. 22, p. 2967, 2020.

[11] D. Avola, L. Cinque, A. Fagioli, G. Foresti, and A. Mecca, "Ultrasound medical imaging techniques: a survey," ACM Computing Surveys (CSUR), vol. 54, no. 3, pp. 1–38, 2021.

[12] A. Lodhi, K. Waite, and I. Alam, "The accuracy of ultrasonography for diagnosis of gallbladder polyps," Radiography, vol. 26, no. 2, pp. e52–e55, 2020.

[13] M. Golla, H. Katiganere Siddaramappa, P. J. Venkata Naga Sai, S. K. Adla, C. Naga, and P. Nijalingappa, "Gbcapsnet: A calibrated capsule network for automated gallbladder disease diagnosis via ultrasound imaging," PloS one, vol. 21, no. 5, p. e0348705, 2026.

[14] M. F. A. Rudro, A. Rahman, Z. Ferdus, N. N. Rahman, S. H. Arman, P. Chowdhury, Z. Al Muhaimin et al., "Gallnet: An explainable deep learning model for multi-class gallbladder disease classification using ultrasound imaging," 2025.

[15] H. R. Nabil, I. Ahmed, A. Das, M. Mridha, M. M. Kabir, and Z. Aung, "Msfe-gallnet-x: a multi-scale feature extraction-based cnn model for gallbladder disease analysis with enhanced explainability," BMC Medical Imaging, vol. 25, no. 1, p. 360, 2025.

[16] S. Jayanthi, I. Kaur, E. L. Lydia, P. R. Babu, J.-C. Yoon, C. Yoon, and G. P. Joshi, "Gallbladder disease diagnosis from ultrasound using squeeze-and-excitation capsule network with convolutional bidirectional long short-term memory," Scientific Reports, 2026.

[17] R. A. Putra, G. A. Pradipta, and P. D. W. Ayu, "Gallbladder disease classification from ultrasound images using cnn feature extraction and machine learning optimization," Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 7, no. 4, pp. 1089–1111, 2025.

[18] M. I. Haque, S. Islam, S. S. Chowa, M. A. H. Rony, K. Fatema, M. M. Hassan, M. R. Islam, M. Z. Hasan, A. Karim, and S. Azam, "Attention-driven deep object detection for improved gallbladder cancer diagnosis from ultrasound images," Engineering Applications of Artificial Intelligence, vol. 176, p. 114757, 2026.

[19] M. Z. Hasan, M. A. H. Rony, S. S. Chowa, M. R. I. Bhuiyan, and A. A. Moustafa, "Gbchv an advanced deep learning anatomy aware model for accurate classification of gallbladder cancer utilizing ultrasound images," Scientific Reports, vol. 15, no. 1, p. 7120, 2025.

[20] J. Li and Y.-Q. Zhou, "Enhanced gallbladder cancer detection via active and self-supervised learning integration: Innovating b-ultrasound image analysis," PLoS One, vol. 20, no. 9, p. e0330781, 2025.

[21] S. Alazwari, J. Alsamri, M. Alamgeer, S. S. Alotaibi, M. Obayya, and A. S. Salama, "Automated gall bladder cancer detection using artificial gorilla troops optimizer with transfer learning on ultrasound images," Scientific Reports, vol. 14, no. 1, p. 21845, 2024.

[22] S. Dadjouy and H. Sajedi, "Gallbladder cancer detection via ultrasound image analysis: An end-to-end hierarchical feature-fused model," IET Image Processing, vol. 19, no. 1, p. e13292, 2025.

[23] S. Basu, M. Gupta, P. Rana, P. Gupta, and C. Arora, "Radformer: Transformers with global–local attention for interpretable and accurate gallbladder cancer detection," Medical Image Analysis, vol. 83, p. 102676, 2023.

[24] S. Basu, M. Gupta, C. Madan, P. Gupta, and C. Arora, "Focusmae: Gallbladder cancer detection from ultrasound videos with focused masked autoencoders," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 11 715–11 725.

[25] T. Kim, Y. H. Choi, J. H. Choi, S. H. Lee, S. Lee, and I. S. Lee, "Gallbladder polyp classification in ultrasound images using an ensemble convolutional neural network model," Journal of Clinical Medicine, vol. 10, no. 16, p. 3585, 2021.

[26] P. Gupta, K. Bose, P. Mudgil, N. Dutta, A. Gulati, G. Prakash, T. D. Yadav, L. Kaman, S. Irrinki, H. Singh et al., "Multiple instance learning approach for automated gallbladder cancer detection using ultrasound imaging: multi-center validation of a deep learning model with the public

dataset contribution," The Lancet Regional Health-Southeast Asia, vol. 47, 2026.
[27] A. M. Obaid, A. Turki, H. Bellaaj, M. Ksantini, A. AlTaee, and A. Alaerjan, "Detection of gallbladder disease types using deep learning: An informative medical method," Diagnostics, vol. 13, no. 10, p. 1744, 2023.
[28] S. Elbedwehy, K. Alnowaiser, A. Saber, E. Hassan, and R. Elmonier, "Explainable hybrid alexnet-efficientvit architecture for accurate gallbladder disease diagnosis," Scientific African, p. e03399, 2026.
[29] N. S. Hassan and A. H. Hamad, "Federated learning for multi-center medical image classification using deep learning models." Ingénierie des Systèmes d'Information, vol. 30, no. 5, 2025.
[30] H. Fujita, T. Wakiya, K. Ishido, N. Kimura, H. Nagase, T. Kanda, M. Matsuzaka, Y. Sasaki, and K. Hakamada, "Differential diagnoses of gallbladder tumors using ct-based deep learning," Annals of Gastroenterological Surgery, vol. 6, no. 6, pp. 823–832, 2022.
[31] S. Rashid, C. J. Das, A. Chauhan, G. Aggarwal, R. C. Joshi, R. Burget, and M. K. Dutta, "Self-attention-guided residual deep neural network with multi-scale dilated feature extraction for automated gallbladder disease diagnosis in ultrasound imaging," Computer Methods and Programs in Biomedicine, p. 109020, 2025.
[32] C. Ge, J. Jang, P. Svrcek, V. Fleming, and Y. H. Kim, "Exploring deep learning applications using ultrasound single view cines in acute gallbladder pathologies: Preliminary results," Academic Radiology, vol. 32, no. 2, pp. 770–775, 2025.
[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: https://jmlr.org/papers/v15/srivastava14a.html
[34] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7
[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2009, pp. 248–255.
[36] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org/
[37] S. Wang, C. Bao, and D. Pei, "Application of data mining technology in the screening for gallbladder stones: A cross-sectional retrospective study of chinese adults," Yonsei medical journal, vol. 65, no. 4, p. 210, 2024.
[38] C.-Y. Yang, H.-L. Kao, Y. C. Chen, C.-F. Kuo, C. H. Liu, and S.-C. Liu, "Automated ct image prescription of the gallbladder using deep learning: Development, evaluation, and health promotion," Acute Medicine & Surgery, vol. 12, no. 1, p. e70049, 2025.
[39] A. Bozdag, M. Yildirim, M. Karaduman, H. B. Mutlu, G. Karaduman, and A. Aksoy, "Detection of gallbladder disease types using a feature engineering-based developed cbir system," Diagnostics, vol. 15, no. 5, p. 552, 2025.
[40] P. Gupta, S. Basu, P. Rana, U. Dutta, R. Soundararajan, D. Kalage, M. Chhabra, S. Singh, T. D. Yadav, V. Gupta et al., "Deep-learning enabled ultrasound based detection of gallbladder cancer in northern india: a prospective diagnostic study," The Lancet Regional Health-Southeast Asia, vol. 24, 2024.
[41] D. Babuji, G. Manikandan, S. Selvaperumal, and D. Venu, "Prediction of gallbladder stone through the use of fuzzy c-means clustering and k-means clustering," in AIP Conference Proceedings, vol. 3161, no. 1. AIP Publishing LLC, 2024, p. 020226.
[42] Y. Zou, L. Wu, C. Zuo, L. Chen, B. Zhou, and H. Zhang, "White blood cell classification network using mobilenetv2 with multiscale feature extraction module and attention mechanism," Biomedical Signal Processing and Control, vol. 99, p. 106820, 2025.

MD. RASHED received the B.Sc. (Engineering) degree in Information and Communication Engineering from Pabna University of Science and Technology (PUST), Pabna, Bangladesh. He is currently a Research Assistant with the Biomedical Robotics, AI & Imaging Network (BioRAIN) Lab. He previously served as an Assistant Researcher with the Department of Information and Communication Engineering, PUST. He received the Best Paper Award at the 2025 IEEE International Conference on Communications, Information, Electronic and Energy Systems (CIEES). His research interests include biomedical image processing, deep learning, machine learning, computer vision, biomedical signal processing, and Internet of Things (IoT).

FAISAL IQBAL HIRON is currently a bachelor's student in the Department of Information and Communication Engineering (ICE) at Pabna University of Science and Technology (PUST), Pabna, Bangladesh. His research interests include deep learning, machine learning, biomedical signal and image processing, computer vision, Internet of Things (IoT), and robotics.

MD. IMRAN HOSSAIN received the Ph.D. degree in Information and Communication Engineering (ICE) from the University of Science and Technology of China in 2022, and the B.Sc. and M.Sc. degrees in ICE from Islamic University, Kushtia, Bangladesh, in 2009 and 2010, respectively. He is currently an Associate Professor with the Department of ICE, Pabna University of Science and Technology, Bangladesh. His current research interests include biomedical signal processing, EEG, ECG, speech separation, and speech enhancement.

MD. SARWAR HOSAIN (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in Information and Communication Engineering (ICE) from the University of Rajshahi, Bangladesh. He began his academic career as a Lecturer with the European University of Bangladesh and subsequently joined Hamdard University Bangladesh and Pabna University of Science and Technology (PUST), where he joined in 2013. He received the M.Phil. degree in 2019 and the Ph.D. degree from Saitama University, Japan, in 2024 under the MEXT Scholarship. His doctoral research focused on speech emotion recognition using bone-conducted speech. He is currently a Professor with the Department of ICE, PUST. His research interests include speech emotion recognition, bone-conducted speech, biomedical image processing, and wireless communication. He is a member of IEEE and ACM and has authored numerous peer-reviewed publications.

OHIDUJJAMAN received the B.Sc. (Honors) and M.Sc. degrees in Computer Science and Engineering from Islamic University, Kushtia, and United International University, Dhaka, Bangladesh, in 2008 and 2014, respectively. He received the Ph.D. degree in Science and Engineering from the Graduate School of Science and Engineering, Saitama University, Japan, in 2024. He was a Research Assistant with the Graduate School of Science and Engineering, Saitama University, Japan, from September 2022 to March 2024. He has also served as a faculty member at Daffodil International University, Hamdard University Bangladesh, and Dhaka International University, and as an Assistant Programmer with the Bangladesh Computer Council (BCC), Agargaon, Dhaka. He received the ICT Scholarship during his B.Sc. studies and the Summa Cum Laude honor during his M.Sc. studies. In 2020, he received the Teaching Practice Excellent Award from Daffodil International University. His research interests include speech signal processing, applications of bone-conducted speech, natural language processing, biomedical signal processing, Universal Networking Language (UNL), and machine learning.

MD. ZAHIDUL ISLAM is a Professor with the Department of Information and Communication Technology at Islamic University, Bangladesh, where he has served since 2003. He is a Senior Member of IEEE (SMIEEE) and has held several leadership positions in the IEEE Computer Society Bangladesh Chapter, including Vice Chair (Activity), Secretary, and Vice Chair (Technical). He is also the Founder and Counselor of the IEEE Student Branch at Islamic University. He received the Ph.D. degree in Computer Engineering from Chonnam National University, South Korea, supported by a research internship at the Samsung Advanced Institute of Technology. He leads the Computer Vision and Intelligent Interfacing Lab (CVIIL), ICT Innovation Lab, and Biomedical Robotics, AI and Imaging Network (BRAIN) Lab. His research interests include computer vision, medical imaging, intelligent robotics, and artificial intelligence-driven healthcare solutions. He holds an international patent on robust particle resampling for visual tracking and has served as Head of the Department of Information and Communication Engineering. He is an Evaluator for BAETE under IEB and a Visiting Research Fellow at Edith Cowan University, Australia. He has authored more than 80 scholarly publications and has extensive experience in research supervision and national and international funded projects.

TETSUYA SHIMAMURA (Senior Member, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees in Electrical Engineering from Keio University, Yokohama, Japan, in 1986, 1988, and 1991, respectively. In 1991, he joined Saitama University, Japan, where he is currently a Professor. He was a Visiting Scholar with Loughborough University, U.K., in 1995 and Queen's University Belfast, U.K., in 1996. He is an author or coauthor of six books and serves on the editorial boards of several international journals. He has also served as a member of organizing and program committees for various international conferences. His research interests include digital signal processing and its applications to speech, image, and communication systems.