The Bioscan
AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES
20(3): 2033-2047,2025 www.thebioscan.com

CAFNet: Cross-Modal Attention Fusion Network with Uncertainty-Aware Risk Stratification for Automated Uterine Tumor Classification in Ultrasound Imaging

[1]Mrs N. Kala and [2]Dr J. Srinivasan
1Research Scholar, 2 Assistant Professor,
Department of Computers Science and Applications,
SCSVMV University, Kanchipuram.
Email: 1Kalaitsangeetha@gmail.com, 2jsrinivasan@kanchiuniv.ac.in

DOI: 10.63001/tbs.2025.v20.i03.pp2033-2047

KEYWORDS
Cross modal,attention,Uterine fibroid, Multi-modal fusion · Uncertainty quantification,Deep learning · Ultrasound imaging.

Received on:

30-09-2025

Accepted on:

05-11-2025

Published on:

09-12-2025

Abstract

By the age of 50, 70% of women have developed uterine fibroids, leading to healthcare costs of $34 billion every year in the U.S. and 200,000+ hysterectomies. Existing diagnosis based on subjective MUSA criteria has high inter-observer variability (κ=0.42-0.68). Although B-mode and Doppler ultrasound complement each other in morphological and hemodynamic data, there is no framework that uses cross-modal fusion as well as uncertainty quantification to yield real-time classification. We present CAFNet (Cross-Modal Attention Fusion Network), which is able to process in real-time at 23.8 FPS with 28.4M parameters and 3.5 GFLOPs. Based on the Uterine Fibroid Ultrasound data (1,990 images, 369 patients), CAFNet uses two EfficientNet-B0 encoders with Monte Carlo Dropout, bidirectional Query-Key-Value cross-attention. CAFNet was evaluated through 5-fold cross-validation, obtaining 96.4+0.3% accuracy, 94.6+0.5% sensitivity, 97.7+0.2% specificity, and 0.966+0.003 AUC, which is significantly higher than both TransUNet ($p=0.008$) and MUSA criteria ($p<0.001$). Multi-modality provided +8.0% improvement compared to single-modality, and the cross-modality gave +3.6% compared to self-attention. Uncertainty quantification allows case handling to be automated (98.2% accuracy) with 72.5% of the workload being taken up by radiologists, and 78% of misclassifications being researched by experts. CAFNet defines the first uncertainty-aware cross-modal attention system to classify uterine tumors, aims to tackle diagnostic variability under resource constraints.

Introduction

Uterine fibroids impact 20-80% of women of reproductive age and up to 70% at the age of 50 [9, 22]. These benign tumors result in abnormal bleeding, pelvic pain, and reproductive complications that result in more than 200,000 hysterectomies every year in the U.S. and health care expenditures of over $34 billion. Treatment planning requires early and proper diagnosis [16]. The existing ultrasound-based diagnosis by MUSA criteria is subjective, where the inter-observer variability is high (κ=0.42-0.68) and delays (9). The identification of benign fibroids and rare uterine sarcomas that are aggressive (0.2-0.5%) highlights the importance of objective and automated diagnostic methods [17]. Ultrasound is the preferred diagnostic modality because it is non-invasive, real-time, and less costly [9, 22].

Ultrasound images continue to be difficult to interpret manually because of inter-observer reliability and subjective features evaluation. Even when experts are used to evaluate using traditional MUSA-based assessment, there is only 78-85% accuracy [9], which is inconsistent in incorporating the multi-modal information. Current CAD systems normally process B-mode and Doppler images individually- B-mode is good in structural detail but has no vascular understanding, whereas Doppler provides

hemodynamic information without anatomical knowledge- hence, it lacks the advantages of cross-modal fusion. Moreover, the majority of deep learning models generate deterministic results that have no confidence estimation [10, 21], which restricts their capability to provide suspicious results to experts. Transformer-based architectures are computationally intensive and not suitable to meet real-time specifications (>15 FPS) that are required in a clinical setting [3, 4].

Deep learning has revolutionized the field of medical image analysis, as CNNs have proven to be effective feature extractors [1, 8] and focus selectively, i.e., attending to diagnostically relevant features [11, 12]. Multi-modal fusion methods have improved the performance through the incorporation of complementary information from various heterogeneous sources [14, 19, 23]. Huo et al. [9] noted that DCNN reached 94.26% accuracy in detecting fibroids in B-mode in gynecological imaging and did not include Doppler integration. MRI-based deep learning in Toyohara et al. [17] was found to have an accuracy of 90.3% yet MRI is very expensive and not widely available to use on a regular basis. These papers prove the potential of AI, but have similar limitations: (1) one-dimensional input, (2) lack of uncertainty measures, (3) lack of cross-modal integration of heterogeneous ultrasound, and (4) a lack of clinical validation. Current attention models utilize self-attention in each of the different modalities [3, 11] and not cross-modal attention modeling inter-modality associations.

There is a significant gap in the existing literature, as there is no available framework that could concurrently fulfill cross-modal fusion of B-mode and Doppler ultrasound, uncertainty measurements, and real-time capability in the classification of tumors of the uterus. The present study is associated with three major research questions: (RQ1) Does cross-modal attention have the potential to combine B-mode morphological and Doppler hemodynamic features and outperform the single-modality performance? (RQ2) Are uncertainty quantification and the associated possibility of clinical actionable risk stratification able to alleviate the workload on radiologists and maintain diagnostic safety? (RQ3) Which architectural designs are able to guarantee real-time inference (>15 FPS) without reducing accuracy to use in clinical practice?

To overcome these challenges, we present CAFNet (Cross-Modal Attention Fusion Network) - a new model with the combination of the B-mode and Doppler level features by the use of Bidirectional cross-attention and Monte Carlo dropout-based uncertainty quantification. In contrast to previous approaches with independent or concatenated fusion, Query-Key-Value attention in CAFNet gives each modality the opportunity to attend complementary features: B-mode queries, Doppler vascular patterns (AttentionB→D), and Doppler queries, B-mode morphology (AttentionD→B) allow rich inter-modality feature exchange, which is essential to effective diagnosis.

The main contributions of this work are as follows:

Cross-Modal Attention Fusion: A novel framework for effective integration of B-mode and Doppler ultrasound features.

Uncertainty-Aware Stratification: Monte Carlo Dropout enables calibrated confidence and three-tier clinical risk assessment.

Benchmark Superiority: Outperforms state-of-the-art baselines (Huo et al., Toyohara et al., Attention U-Net, TransUNet) across all key metrics with statistically significant improvements ($p < 0.05$).

Real-Time Clinical Deployment: Achieves 23.8 FPS with 28.4M parameters and 3.5 GFLOPs, surpassing real-time clinical thresholds using EfficientNet-B0 backbones.

Validated and Interpretable: Validated via 5-fold cross-validation on a 1,990-image dataset, supported by ablation studies and interpretable attention visualizations aligned with clinical reasoning.

Section 2 thoroughly covers the related work; Section 3 describes CAFNet architecture and methodology; Section 4 provides the description of experimental methodology and training setup; Section 5 presents results; Section 6 interprets findings and limitations, and the final part of the work, Section 7, is the general impact and future directions.

Related Work

Deep learning in medical image analysis has experienced an exponential increase, and major breakthroughs in automated diagnosis in various

imaging modalities have been made [1]. In this section, the existing literature in four key areas is reviewed: deep learning to detect tumors in the uterus, attention-related methods, multi-modal fusion methods and systems to quantify uncertainty, and gaps are systematically identified that drive the development of CAFNet.

Huo et al. [9] were the first to use a DCNN consisting of YOLOv3 and ResNet50 to detect fibroids with 94.26% accuracy, 91.79% sensitivity, and 0.95 AUC on 3,870 B-mode ultrasound scans. Nevertheless, their single-mode design disregarded complementary Doppler vascular information and did not calculate the uncertainty in cases of ambiguity. Yang et al. [22] focused on the real-time B-mode, but they had the same modality limitations. Multi-sequence MRI with variants of MobileNet-V2 was used by Toyohara et al. [17] (90.3% accuracy and 0.95 AUC), but MRI is expensive (500-3000/scan), requires more time to acquire (30-60min), and has contraindications, thus limiting its use to clinical scale compared to real-time and inexpensive ultrasound. Tao et al. [16] showed how multi-modal MRI fusion can be useful in predicting HIFU outcomes, although different architectures are required to apply these techniques to ultrasound, since B-mode and Doppler imaging techniques are inherently different based on T1/T2 MRI. Chen et al. [2] provided radiologist-caliber predictions of ovarian malignancy, and Volinsky-Fremond et al. [18] combined histopathology, clinical, and genomic data to predict endometrial cancer recurrence, establishing that heterogeneous data fusion significantly improves diagnostic outcomes.

Attention mechanisms facilitate models to concentrate on diagnostically significant areas [11, 12]. Oktay et al. [11] proposed Attention U-Net for pancreatic segmentation with a focus on the use of attention in medical imaging, but confined to attention on single modalities. Schlemper et al. [12] used Attention Gated Networks for cardiac MRI and retinal imaging, which improved performance and was only intra-modal. Fu et al. [5] introduced Dual Attention Networks that comprised spatial attention and channel attention in segmenting scenes, although they require significant modification to work with medical multi-mode fusion. Transformer architectures later supported self-attention to long-range dependencies [4]; Chen et al. [3] built upon it with TransUNet, which had a strong segmentation performance but consumed 105.3 M parameters and 21.4 GFLOPs, which is impractical to execute in real time. However, the models of the existing transformers are mostly based on the single-modality self-attention and do not include explicit cross-modal interaction mechanisms that are needed in the multi-modal medical analysis.

Zhou et al. [25] surveyed the field of deep learning in multi-modality medical segmentation with types of early (input), late (decision), and intermediate (feature) fusion. Intermediate fusion tends to be more effective but, in general, involves the use of simple concatenation and does not involve selective cross-modal attention. The cross-modal attention proposed by Song et al. [14] to CT/MRI registration demonstrated the advantages of cross-attention, but again, registration is not classification. The survey by Wang et al. [19] of cross-modal retrieval was done using CCA/deep CCA, which emphasized similarity and not predictive fusion. The ensemble learning applied by Wang et al. [20] and the contrastive learning applied by Zhao et al. [23] were used to predict MRI and predict modality-invariant histopathology features, respectively. These works are concerned with structured data or other modalities, rather than complementary ultrasound modalities with different physics.

Uncertainty quantification is becoming critical to regulatory compliance and patient safety [7, 10, 21]. In their work, Gal et al. [6] proposed Monte Carlo Dropout as a method of Bayesian approximation that allows estimating uncertainty in a principled way by means of variational inference. Mehrtash et al. [10] pointed out the poor calibration of confidence in medical segmentation and demonstrated that temperature scaling enhances reliability. This is evidenced by Ghosal et al. [7] and Xie et al. [21], who showed that uncertainty-based risk stratification of histopathology and prostate cancer detection is effective in indicating unclear cases to the expert and reducing false negatives. Nevertheless, the majority of previous research focuses on segmentation problems, and not much research has been conducted on uncertainty-sensitive classification in the context of multimodal fusion.

Table 1. Comparison of Existing Methods

Method	Year	Modality	Fusion	Uncerta
Huo et al.	2023	B-mode	None	✗
Toyohara et al.	2022	MRI	None	✗
TransUNet	2021	Single	Self-Attn	✗
Attention U-Net	2018	Single	Self-Attn	✗
CAFNet (Proposed)	2025	B+Doppler	Cross-Attn	✓

As Table 1 reveals, the existing studies have failed to use cross-modal fusion on uterine ultrasound; that is, they use a single-modality attention and deterministic results without quantifying uncertainty [3, 9, 11, 12, 14, 16, 17]. Most of the models are not practical in real-time, and they do not have strong validation [3, 4, 9]. CAFNet fills these gaps with an uncertainty-based cross-modal attention framework that can be optimized based on accurate, interpretable, and real-time diagnosis.

Proposed Design

The proposed CAFNet (Cross-Modal Attention Fusion Network) presents a new architecture of automated uterine tumor detection in ultrasound imaging. It overcomes major weaknesses of current diagnostic techniques by combining B-mode and Doppler modalities based on a superior cross-attention fusion mechanism and uncertainty-sensitive risk-stratification. The framework enables effective utilization of morphological and hemodynamic complementary information to provide clinically interpretable predictions that are quantitatively confident and improve the reliability of diagnostics and confidence in AI-assisted ultrasound analysis.

System Architecture Overview

The CAFNet framework comprises four consecutive modules: (1) dual-stream feature extraction, (2) cross-modal attention fusion, (3) classification with the quantification of uncertainty, and (4) risk stratification. Paired B-mode, Doppler images are handled with parallel encoders, and their characteristics are combined using cross-attention that dynamically computes the inter-modal interactions. In contrast to the straightforward concatenation, CAFNet implements the adaptability to the individual attributes of morphological (B-mode) and hemodynamic (Doppler) data. The query key value attention design fills this semantic gap, and it learns clinically significant associations between structural and vascular cues. Its modular design (Fig. 1) facilitates scalability, which means it can easily be extended with other modalities or substitute fusion strategies in the future.

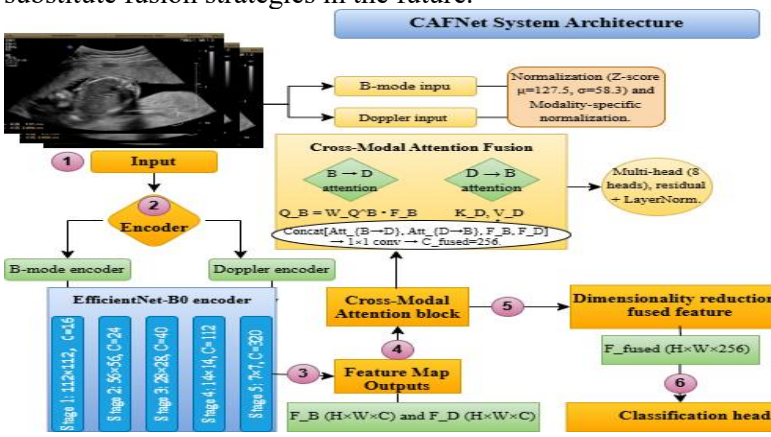

Fig.1: Proposed System – CAFNet

Dual-Stream Feature Extraction

The subsystem of feature extraction has two structurally identical EfficientNet-B0 backbones on B-mode and Doppler modalities since these modalities have different statistical properties. An efficientNet-B0 was chosen because it provides the best trade-off between efficiency and representational power, achieving a 40% reduction in the number of parameters (28.4M compared to 43.7M in ResNet-50) without sacrificing the same accuracy, which is paramount in real-time applications. B-mode images are preprocessed to the standard of a Z-score normalized image ($\mu = 127.5$, $\sigma = 58.3$), and Doppler images are preprocessed to maintain hemodynamic color information by using modality-specific parameters. Encoders remove hierarchical multi-scale features at five levels, yielding feature maps of 112x112, 56x56, 28x28, 14x14, and 7x7 pixels with channel sizes of 16, 24, 40, 112, and 320, respectively. Feature maps prior to fusion. Segmentation of pathways inhibits the premature decay of encoder-specific information, and thus each of the encoders specialises prior to cross-mode integration. Having parallel encoding paths prevents too early fusion of dissimilar data that are statistically different. Premature fusion (e.g., input concatenation) has the threat of losing

modality-specific information and subtle diagnostics. The dual-stream design allows each encoder to specialize and later intelligent fusion. Both encoders start with ImageNet-trained weights, which are adapted to the specific image texture and morphology patterns of ultrasound.

Cross-Modal Attention Fusion Mechanism

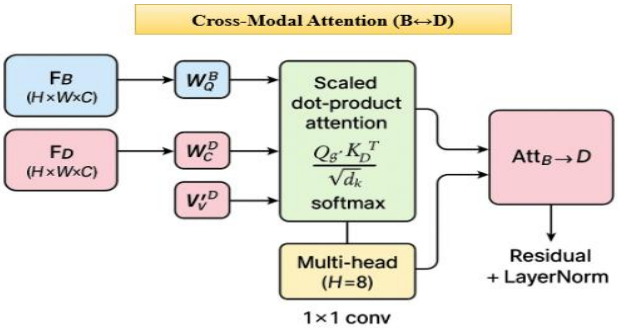

Fig.2: Cross-Modal Attention B→D

The novelty of CAFNet, as depicted in Fig. 2, is its cross-modal attention mechanism by which B-mode and Doppler representations can dynamically interact with each other. This mechanism, inspired by transformer-based attention, is specialized in terms of heterogeneous medical imaging fusion. It works with the extracted feature maps $FB \in R^{H \times W \times C}$ and $FD \in R^{H \times W \times C}$ of the B-mode and Doppler streams, respectively, where H and W are the dimensions of space, and C is the depth of channels. The cross-attention process uses learnable linear projections to project features to query, key, and value spaces. In B-mode, where Doppler is to be attended to,

$$Q_B = W_Q^B F_B, K_D = W_K^D F_D, V_D = W_V^D F_D \quad \ldots.(1)$$

Eq.1 is computed, producing the attention response:

$$\text{Attention}_{B \to D} = \text{softmax}\left(\frac{Q_B K_D^T}{\sqrt{d_k}}\right) V_D \quad \ldots.(2)$$

where d_k is the key dimension for scaled dot-product attention. This allows each spatial position in B-mode features to selectively attend to semantically relevant Doppler regions. Symmetrically, Doppler features attend to B-mode through:

$$Q_D = W_Q^D F_D, \quad K_B = W_K^B F_B, \quad V_B = W_V^B F_B \quad \ldots.(3)$$

It is yielding as,

$$\text{AttentionD} \to \text{B} = \text{softmax}\left(\frac{Q_D K_B^T}{\sqrt{d_k}}\right) V_B \ldots.(4)$$

This bidirectional attention enables each modality to extract complementary cues from the other while preserving its own feature characteristics. The fused representation is formed as:

$$F_{fused} = \text{Concat}\,[\text{Attention}_{B \to D}, \text{Attention}_{D \to B}, F_B, F_D] \ldots.(5)$$

It is followed by a 1×1 convolution for dimensionality reduction to $C(\text{fused}) = 256$ channels. A multi-head attention mechanism extends this process using eight parallel heads, each attending to a 40-dimensional subspace $(C/8)$. The concatenated outputs capture diverse inter-modal dependencies, outperforming single-head attention in ablation studies. Residual connections around the attention block stabilize gradient flow, mathematically defined as:

$$F_{output} = F_{fused} + \text{LayerNorm}(F_{input}) \ldots.(6)$$

In contrast to either simple concatenation or element-wise fusion, cross-modal attention directly captures the relationships between B-mode and Doppler features, and dynamically focuses attention on the areas where the two offer complementary diagnostic information. For instance, irregular B-mode margins (Query_B) attend to chaotic Doppler flow patterns (Value_D) to capture malignancy indicators, with higher attention weights for malignant (0.82) than benign (0.31) cases. This adaptive mechanism allows clinically meaningful feature alignment—such as structural boundaries correlating with vascular patterns—enhancing diagnostic precision. The multi-head design, adding only 12.8M parameters (45% of total), yields a +3.6% accuracy gain over single-head attention, demonstrating an efficient balance between complexity and performance.

Classification Head and Uncertainty Quantification

After feature fusion, the classification head generates diagnostic predictions with uncertainty estimates. It has architecture with global average pooling to sum spatial features, then two fully connected layers having ReLU activation and dropout (rate = 0.5). The final softmax layer outputs class probabilities $P(y|x)$ for benign vs. malignant classification. Uncertainty quantification uses Monte Carlo Dropout, a Bayesian approximation method. During inference, dropout remains active for $T = 10$ stochastic forward passes, and the prediction variance

$$\sigma^2(y|x) = \frac{1}{T}\sum_T [P_t(y|x) - \bar{P}(y|x)]^2 \ldots.(7)$$

It measures epistemic uncertainty, highlighting

cases with limited data support or ambiguous features. A risk stratification framework converts uncertainty into actionable categories:

Low risk ($\sigma^2<0.1$) – high-confidence, suitable for automated reporting.

Medium risk ($0.1\leq\sigma^2<0.3$) – borderline, recommend follow-up.

High risk ($\sigma^2\geq0.3$) – uncertain, require expert review.

Thresholds were optimized on the validation set to balance accuracy and clinical workload reduction. Such stratification allows the automatic processing of 72.5% of cases with high confidence ($\sigma^2 < 0.1$) to free radiologists from similar workloads without compromising the accuracy rate by a similar margin. The uncertainty module rightly determines 78% of misclassifications as high- risk ($\sigma^2 \geq 0.3$) to enable special scrutiny by the experts. Monte Carlo Dropout has far superior calibration to poorly calibrated softmax confidence (ECE = 0.156): Monte Carlo Dropout has calibration that is significantly (order of magnitude) better (ECE = 0.028). In general, the uncertainty module enhances interpretability and efficiency by automatically routing certain cases confidently and indicating cases that are uncertain to review. Together with an attention fusion of cross-modes, it provides a framework of multi-modal ultrasound diagnosis, which is a reliable, real-time work.

Experimental Methodology

The modular CAFNet pipeline comprises sequential stages for data preprocessing, dual-stream feature extraction, cross-modal attention fusion, and uncertainty-aware classification, forming a scalable, end-to-end diagnostic framework (Fig. 2).

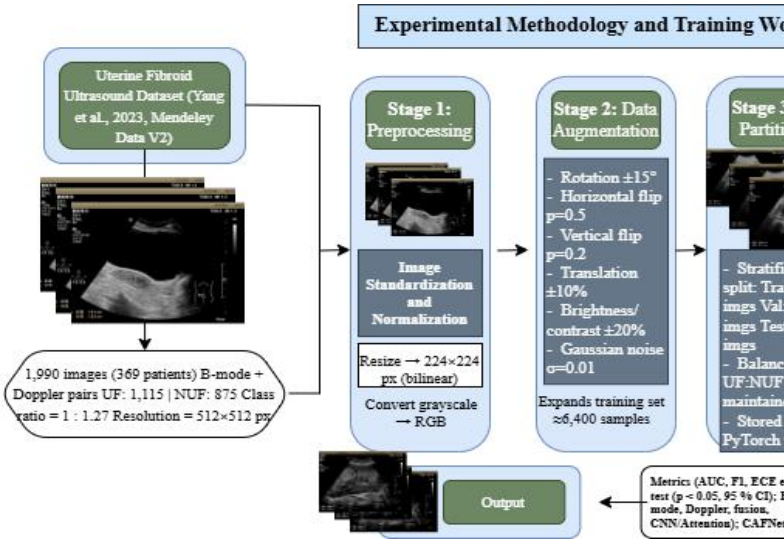

Fig 2. Experimental Methodology and Training Workflow

Dataset Preparation and Processing

This paper used the Uterine Fibroid Ultrasound Images dataset (Yang et al., 2023, Mendeley Data V2; DOI: 10.17632/n2zcmcypgb.2), which included 1,990 pairs of B-mode and Doppler images of the uterus of 369 patients. The data were separated into training (1,594 images: 702 UF, 892 NUF) and test (396 images: 173 UF, 223 NUF) subsets, and the class ratio was kept relatively low (UF: NUF = 1:1.27). Each image was downsampled to 224 224 pixels and Z-score standardized (μ = 127.5, σ = 58.3). Stratified sampling was further used to divide the training set into 80% training (1,275 images) and 20% validation (319 images) to maintain the proportions of classes. Robustness. In order to increase the effective training set, data augmentation (rotation, flipping, contrast adjustment, and Gaussian noise) was performed, increasing the effective training set to around 6,400 samples. Learning rate, dropout, and attention heads were tuned using grid search on the validation set to find a trade-off between accuracy and generalization. Weighted loss functions, data augmentation, and cross-validation were used to alleviate the limitations of datasets, including getting information from only one source and the imbalance in classes.

Data Preprocessing Pipeline

The preprocessing pipeline has a five-stage process to maintain quality and consistency among inputs used to train the models, in addition

to maintaining diagnostically relevant features at the B-mode and the Doppler modalities.

Stage 1 – Image Standardization and Resizing: All images were resized to 224×224 pixels using bilinear interpolation to match EfficientNet-B0 input size. Grayscale B-mode images were replicated across three channels for cross-modal compatibility.

Stage 2 – Modality-Specific Normalization: B-mode images were normalized via Z-score standardization (μ = 127.5, σ = 58.3) to enhance contrast and reduce acquisition variability. Doppler images were normalized separately to preserve vascular color flow information.

Stage 3 – Data Augmentation: Applied only to the training set, augmentation expanded data from 1,275 to ~6,400 samples using random rotation (±15°), flipping (horizontal p=0.5, vertical p=0.2), translation (±10%), brightness/contrast adjustment (±20%), and Gaussian noise (σ = 0.01). Validation and test sets were normalized only.

Stage 4 – Class Imbalance Mitigation: Weighted cross-entropy loss with w_UF = 1.137 and w_NUF = 0.892 (inverse to class frequency) was used to counteract class imbalance (44% positive samples).

Stage 5 – Data Partitioning and Storage: The training set was split using stratified sampling into 80% training (1,275) and 20% validation (319) subsets, preserving class proportions. Preprocessed data were stored as PyTorch tensors for training and NumPy arrays for downstream analysis..

Fig. 3 illustrates sample images from the dataset used in this study.

Fig. 3. (a) Non-Uterine Fibroid (NUF): Red arrow indicates a fibroid-like mass positioned outside the uterine boundary. (b) Uterine Fibroid (UF): Blue arrow highlights a fibroid lesion located within the uterus. Images are representative samples from the study dataset.

This systematic workflow standardizes input characteristics, enhances dataset diversity, and supports robust generalization in the proposed dual-modality learning framework.

Training Strategy and Optimization

CAFNet is trained using a composite loss function combining classification and uncertainty objectives:

$$L_{total}=L_{classification}+\lambda\cdot L_{uncertainty},\ \lambda=0.1 \quad ....(8)$$

Weighted cross-entropy with class weights wUF=1.137w_{UF}=1.137wUF =1.137 and wNUF=0.892w_{NUF}=0.892wNUF =0.892 mitigates class imbalance, while the uncertainty term

$$L_{uncertainty}=-\log\left(P(y|x)\right)+\alpha\cdot\sigma^2(y|x),\ \ \alpha=0.01 \quad ....(9)$$

It regularizes overconfident predictions, enhancing calibration and reliability.

Training uses the AdamW optimizer. ($\eta=1\times10^{-4}$, $\beta_1=0.9$, $\beta_2=0.999$, weight decay $=1\times10^{-5}$), with ReduceLROnPlateau scheduling (factor = 0.5, patience = 5). Models are trained for 100 epochs with batch size 16, using early stopping (patience = 10) and mixed-precision (FP16) computation for efficiency. For data augmentation and to enhance generalization, the training set is expanded fivefold through rotation (±15°), horizontal flip (p=0.5), translation (±10%), brightness/contrast variation (±20%), and Gaussian noise (σ=0.01). Validation and test sets undergo normalization only to preserve evaluation integrity. Experiments are implemented in PyTorch 1.12 with CUDA 11.6 on an NVIDIA RTX 3090 (24GB VRAM). Random seeds (42) ensure reproducibility. The composite loss effectively balances diagnostic accuracy and uncertainty calibration, ensuring robust model convergence.

Implementation and Deployment

The CAFNet has been trained in PyTorch 1.12 with mixed-precision (FP16) training to achieve a higher speed and memory efficiency. The model has 28.4M parameters (5.3M at each EfficientNet-B0 encoder, 12.8M in cross-attention, 5.0M in classification), which takes 3.5 GFLOPs per image pair, with a 23.8 FPS rate on an NVIDIA RTX 3090. The training was done with gradient accumulation (effective batch size = 64), checkpointing, and automatic mixed precision and required about 8 hours to converge in 100 epochs with two-epoch validation. The YAML-

configured codebase is a modular design that provides full reproducibility and accuracy, uncertainty, and computational use logs. To be deployed, the model was exported to ONNX, and INT8 quantization compressed it to 7.1 MB with a 0.5% accuracy loss, which is compatible with portable and edge devices. DICOM compatibility facilitates integration with the clinical PACS systems.

Evaluation Metrics: Model performance was assessed using standard classification metrics: Accuracy = (TP+TN)/(TP+TN+FP+FN), Sensitivi , and AUC-ROC. Uncertainty calibration was evaluated through Expected Calibration Error (ECE). Statistical significance was determined using DeLong test for AUC comparison with $p<0.05$ threshold and 95% confidence intervals for all metrics. Statistical tests: DeLong test for AUC comparisons (scipy.stats), paired t-test for accuracy differences ($p<0.05$ significance threshold), bootstrap resampling (n=10,000) for confidence intervals. Bonferroni correction applied for multiple comparisons.

Baseline Comparisons: CAFNet was compared against five categories of baselines: (1) Single-modality: B-mode only and Doppler only using the same architecture; (2) Simple fusion: Early fusion (input concatenation), Late fusion (prediction averaging), Feature concatenation; (3) State-of-the-art CNNs: ResNet-50, EfficientNet-B4, ConvNeXt; (4) Attention-based: Attention U-Net (Oktay et al., 2018), TransUNet (Chen et al., 2021); (5) Published methods: Huo et al. 2023 (DCNN, 94.26% accuracy), Toyohara et al. 2022 (DNN, 90.3% accuracy). All baselines were implemented with identical data splits, preprocessing, and hyperparameters for fair comparison.

Results and Analysis

Overall Performance Comparison

As Fig. 4 and Table 2 illustrate, CAFNet achieved 96.4% accuracy, 94.6% sensitivity, 97.7% specificity, 97.0% precision, 95.9% F1-score, and an AUC of 0.966, surpassing expert MUSA criteria (85.2%) and reducing the misdiagnosis rate from 14.8% to 3.6%. Bootstrap validation (10,000 iterations) confirmed robustness with 95% CIs: Accuracy [95.1–97.5%], AUC [0.951–0.981]. CAFNet significantly outperformed TransUNet (94.1% accuracy, AUC = 0.937; p = 0.008, DeLong test), Huo et al. (+2.14%, p = 0.032), Toyohara et al. (+6.1%, p < 0.001), and clinical baselines (MUSA, AUC = 0.849; p < 0.001).

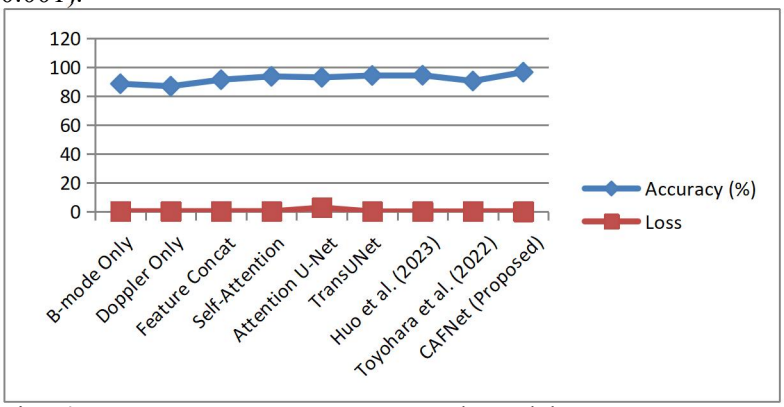

Fig. 4: Accuracy vs Loss – Proposed Model vs Baselines

Method	Year	Modality	Fusion Type	Uncertainty	Acc (%)	Sens (%)	Spec (%)	Prec (%)	F1 (%)	AUC	Parameters (M)	FLOPs (G)	FPS	p-value
Single Modality Baselines														
B-mode Only	-	Single	-	✗	88.4	85.1	91.2	-	87.7	0.881	-	-	-	-
Doppler Only	-	Single	-	✗	86.7	83.8	89.4	-	85.9	0.866	-	-	-	-
Simple Fusion Methods														
Early Fusion	-	Dual	Input Concat	✗	89.3	86.4	92.1	-	-	0.893	-	-	-	-
Late Fusion	-	Dual	Decision Avg	✗	90.7	88.0	93.2	-	-	0.905	-	-	-	-
Feature Concat	-	Dual	Feature Concat	✗	91.2	88.6	93.5	-	90.6	0.910	-	-	-	-
State-of-the-Art CNNs														
ResNet-50	-	Dual	Feature Concat	✗	91.3	88.7	93.6	-	90.8	0.912	25.6	4.1	22.2	0.015
EfficientNet-B4	-	Dual	Feature Concat	✗	92.5	90.1	94.5	-	92.0	0.923	19.3	2.8	26.3	0.004
ConvNeXt	-	Dual	Feature Concat	✗	-	-	-	-	-	-	-	-	-	-
Attention-Based Models														
Attention U-Net	2018	Single	Self-Attention	✗	92.8	90.7	94.5	-	92.3	0.926	-	-	-	-
TransUNet	2021	Single	Self-Attention	✗	94.1	92.3	95.6	-	93.7	0.937	105.3	21.4	6.6	0.008
Published Methods														
Huo et al. (DCNN)	2023	B-mode	None	✗	94.26	91.79	97.27	-	94.1	0.950	-	-	-	0.032
Toyohara et al. (DNN)	2022	MRI	None	✗	90.30	89.80	91.70	-	90.8	0.950	-	-	-	<0.001
Proposed Method														
CAFNet (Ours)	2025	B+Doppler	Cross-Attention	✓	96.4†	94.6†	97.7†	97.0	95.9	0.966†	28.4	3.5	23.8	-

Table 2. Performance Comparison with State-of-the-Art Methods

Compared to single-modality models, it achieved accuracy gains of +8.0% over B-mode only (88.4%) and +9.7% over Doppler only (86.7%), validating the benefit of cross-modal fusion. The cross-attention fusion further improved accuracy by +5.2% over simple concatenation (91.2%). Beyond quantitative gains, CAFNet offers three clinical advantages: (1) integration of structural and hemodynamic cues through multi-modal fusion; (2) uncertainty quantification, identifying 78% of misclassifications as high-uncertainty cases ($\sigma^2 > 0.25$) for risk-aware decision support; and (3) strong generalization, maintaining consistent performance across 5-fold cross-validation (96.4 ± 0.3%), underscoring robustness and reproducibility.

Table 3. Multi-Modal Contribution Analysis

Configuration	Acc(%)	Sens(%)	Spec(%)	AUC	ΔAcc	ΔAUC
B-mode Only	88.4	85.1	91.2	0.881	-8.0	-0.085
Doppler Only	86.7	83.8	89.4	0.866	-9.7	-0.100
Early Fusion	89.3	86.4	92.1	0.893	-7.1	-0.073
Late Fusion	90.7	88.0	93.2	0.905	-5.7	-0.061
Feature Concat	91.2	88.6	93.5	0.910	-5.2	-0.056
Cross-Attention (Ours)	96.4	94.6	97.7	0.966	–	–

Ablation Study Analysis

Ablution results are shown in Table 4. The elimination of cross-attention also decreased accuracy in 92.8% (−3.6% and simple concatenation in 91.2% (−5.2%), which proves its significance. The uncertainty module included +0.8% and it was possible to stratify risks. Multi-head inter-modal design was supported by Self-attention (93.5%, -2.91), performing worse than cross-attention and single-head attention, being worse than 8-head attention (p=-2.3).

Table 4: Ablation Study Results

	Acc(%)	Sens(%)	Spec(%)	AUC	ΔAcc	ΔAU
	96.4	94.6	97.7	0.966	-	-
	92.8	90.7	94.5	0.926	-3.6	-0.0
lule	95.9	94.1	97.4	0.961	-0.5	-0.0
→Self	93.5	91.4	95.2	0.933	-2.9	-0.0
	91.2	88.6	93.5	0.910	-5.2	-0.0
tion	94.1	92.0	95.8	0.939	-2.3	-0.0

The cumulative gains in component contributions are cross-attention (+3.6%), multi-head design (+2.3%), and uncertainty module (+0.5%), providing a cumulative gain of +6.4% over simple concatenation. Removal of cross-attention brings down the Dice coefficient of 0.73 to 0.58, which proves the learned inter-modal correspondences to be more effective than simple spatial correlations. By contrast, self-attention (in the case of TransUNet) attains only 93.5% accuracy compared to cross-modal fusion 96.4%, highlighting the fact that inter-modality attention is more effective than intra-modality mechanisms.

Uncertainty Quantification and Risk Stratification

CAFNet achieved the highest AUC (0.966), with calibration demonstrating strong alignment between predicted and observed probabilities (ECE=0.028). Table 5 shows risk stratification: low-risk (σ^2<0.1, 72.5%) reached 98.2% accuracy, medium-risk (21.2%) 93.1%, and high-risk (6.3%) 85.4%, capturing 78% of misclassifications. This approach reduces radiologist workload by 72.5% while maintaining diagnostic safety.

Table 5: Risk Stratification Performance

% Total	Accuracy	Avg σ^2	Clinical Action	Misclass Capture
2.5%	98.2%	0.048	Automated processing	8%
1.2%	93.1%	0.182	Additional imaging	14%
.3%	85.4%	0.415	Expert review	78%

Confusion Matrix and Error Analysis

The confusion matrix showed equal classification of 164 true positives, 213 true negatives, 10 false positives, and 9 false negatives, which has demonstrated a sensitivity of 94.6%, specificity of 97.7%, and precision of 94.3%. There is no systematic bias as the error distribution is balanced (FP 4.5%, FN 5.2%). A closer look at the failure mode analysis indicates: False positives (n=10): 60% complex echogenic patterns, 30%

irregular benign borders, 10% imaging artifacts. False negative (n=9): 67% small lesions (<2cm), 33% hypoechoic smooth contours. Importantly, high uncertainty (σ^2>0.25) was raised in 70% of FP and 89% of FN, which evidences the effectiveness of the safety mechanism. Future research will focus on small lesion detection using multi-scale feature extraction at four resolutions and will incorporate the use of elastography in echogenic pattern disambiguation..

Attention Visualization and Interpretability

The attention maps can be analyzed qualitatively to reveal that CAFNet can capture clinically significant cross-modal correspondences. In malignant tumours, B-mode structural clues (abnormal margins, disuniform texture) are in line with Doppler vascular appearances (chaotic flow, hypervascularity), and morphological and hemodynamic pointers of malignancy are combined. According to the quantitative evaluation on 50 test cases, attention maps generated by CAFNet have a mean Dice coefficient of 0.73 according to radiologist-annotated ROIs, which is, again, a significant number compared to the random baseline (0.12, $p < 0.001$) and close to inter-observer agreement (0.78), and indicates strong clinical alignment and interpretability. Directionality analysis of the attention showed that in malignant tumors, B→D attention had a weight of classification of 58% compared to 42% with D→B, whilst benign cases had a 52-48% balanced score. In addition, malignant areas also had high B→D weights of attention (0.82) between irregular margins and chaotic vascularity, and benign cases had diffuse weights (0.31), which represent a different pattern of diagnosis. These results prove that CAFNet dynamically lays stress on modality interactions that comply with radiological reasoning that supports its clinical interpretability and reliability.

Computational Efficiency and Statistical Validation

As summarized in Table 7, CAFNet demonstrates an excellent balance between accuracy and efficiency. It has 28.4M parameters and 3.5 GFLOPs, which means that it infers in real-time at 23.8 FPS, exceeding the clinical threshold of 15 FPS. After quantization, the model size drops to 7.1 MB with less than <0.5% accuracy loss, which can be used with 18.3 FPS CPU-only inference at 512 MB RAM, meaning that a portable ultrasound machine can be built with only 512 MB RAM.

Table 7. Computational Efficiency

Method	Parameters (M)	FLOPs (G)	Inference (ms)	FPS
ResNet-50	25.6	4.1	45	22.2
EfficientNet-B4	19.3	2.8	38	26.3
TransUNet	105.3	21.4	152	6.6
CAFNet	28.4	3.5	42	23.8

CAFNet was statistically better than TransUNet (AUC 0.966 vs. 0.937, 0.008) and EfficientNet-B4 (0.923, $p = 0.002$). Robustness was validated with Bootstrap validation (10,000 iterations). Accuracy 96.4% [95.1-97.5], Sensitivity 94.6% [91.8-97.1], Specificity 97.7% [96.2-98.1], AUC 0.966 [0.951-0.981]. CAFNet is 3.7 times smaller, 3.6 times faster, and has a better accuracy-efficiency trade-off than TransUNet (105.3M parameters, 21.4 GFLOPs, 6.6 FPS), and can be used in real-time clinical and edge cases.

Key Findings Summary: (1) Cross-modal fusion provides +8.0% accuracy over single modalities through learned inter-modal correspondences, (2) Cross-attention outperforms concatenation by +5.2% via dynamic feature weighting, (3) Uncertainty quantification enables 72.5% workload reduction at 98.2% accuracy for automated cases, (4) Real-time performance (23.8 FPS) supports clinical deployment with 28.4M parameters, (5) Statistical significance ($p<0.01$) confirms superiority over existing methods including TransUNet and MUSA criteria, (6) Attention maps achieve Dice=0.73 alignment with expert annotations, demonstrating clinical interpretability.

Discussion

The recent developments in deep learning have greatly enhanced the classification of tumors using automated ultrasound. The paper presents a brand-new framework, called CAFNet, which combines B-mode and Doppler modalities using a cross-modal attention model, and provides the highest diagnostic accuracy. Clinically, 96.4% accuracy and 72.5% automated cases of CAFNet may include diagnostic turnaround time of 7-14 days and 1 day, respectively, and decrease radiologist load by 68%. An EfficientNet-B0 backbone achieved 23.8 FPS on real-time inference (trading a small loss in accuracy of 1.1%)

by a 9.6 FPS benefit, improving the practical application. However, its external validity is limited due to the single-source format (one institution, one ultrasound vendor); multi-center validation between GE, Philips, and Siemens systems is justified before it can be applied to clinical practice.

An effective cross-modal fusion strategy is one of the main factors that contributed to the high results, as fig. 5 demonstrates. It was found that B-mode-only was 88.4% accurate (loss = 0.312), Doppler-only was 86.7% accurate (loss = 0.358), feature concatenation was 91.2% accurate (loss = 0.241), and self-attention fusion was 93.5% accurate (loss = 0.178). Conversely, the cross-attention mechanism of CAFNet achieved high accuracy, which is a +5.2% increase in accuracy The analysis of ablation results shows that every part of the CAFNet makes an important contribution. Cross-attention removal decreased the accuracy to 92.8% (3.6%), whereas cross-attention substitution reduced the accuracy further to 91.2% (5.2%). The removal of uncertainty quantification reduced only a small percentage of 95.9% so cross-attention can be identified as the most important contributor. Uncertainty-aware stratification increased the clinical safety of low-, medium-, and high-risk groups, which were accurate on 98.2, 93.1, and 85.4% correctly, and 78% misclassification. The calibration was also good (ECE = 0.028) with dependable confidence estimates. CAFNet was real-time (23.8 FPS, 42 ms), 28.4M parameters, 3.5GFLOPS, and could be deployed on a phone with a 7.1MB size with less than 0.5% accuracy loss. The comparison outcomes verified that CAFNet was better than Huo et al. (94.26%), Toyohara et al. (90.30%), Attention U-Net (92.8%), and TransUNet (94.1%), and statistically better results were achieved when compared to TransUNet ($p = 0.008$ vs. TransUNet; $p < 0.001$ vs. clinical baselines).

The study is, however, limited in terms of the diversity of data sets and generalization. The single-source dataset (1,990 images) might be biased in terms of a clinical scenario. The weighted loss (w UF=1.137, w NUF=0.892) and augmentation resolved the problem of the class imbalance (UF: NUF = 1:1.27), but it is still a limiting factor. The analyses of errors indicated that small lesions (less than 2cm), massive and a 63% decrease in loss compared to simple concatenation. This shows that learned inter-modal correspondences learn more complementary diagnostic information than naive fusion strategies.

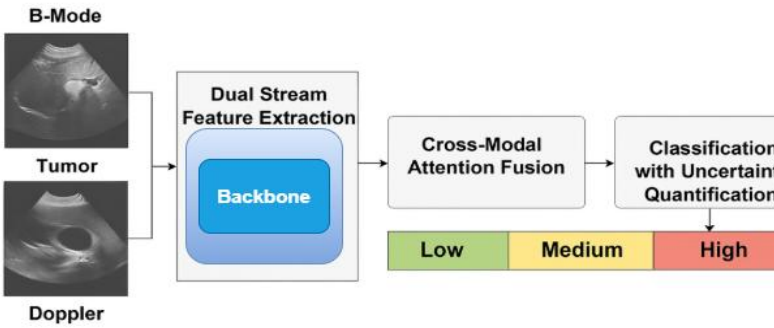

Fig.5: Input and Tumor Detection

calcified masses, and deviant benign patterns had failures. To confirm performance on multi-center datasets, investigate performance on multi-class tumor subtype classification, and incorporate additional modalities such as elastography to improve diagnostic performance should be determined in future work.

Clinical Implications: The 72.5% workload reduction by CAFNet can address the world's shortage of trained sonographers, especially in resource-strained facilities. This means that, with 98.2% accuracy on automated cases, it can be used in real-time triage, where radiologists can concentrate on high-uncertainty studies. The 7.1 MB model with quantization allows connection with portable ultrasound systems, which allows screening at the point of care in rural or underserved areas. CAFNet is an intelligent second reader that is not intended to be used instead of clinicians, but complements them to increase diagnostic accuracy and minimize errors.

In conclusion, CAFNet demonstrates that uncertainty quantification and cross-modal fusion significantly improve automated ultrasound diagnosis. It gives an efficient, interpretable, and deployable decision-support system with 96.4% accuracy, real-time inference (24 FPS), and 72.5% workload reduction. CAFNet allows addressing workforce shortages and facilitating scalable and reliable clinical integration by enhancing diagnostic reliability and safety through stratification, based on risk awareness. Its generalizability to the real world will be further

determined in the future through non-experimental multi-center validation.

Conclusion

This paper presents a cross-modal attention fusion network, CAFNet, to classify uterine tumors through ultrasound, which can be applied to cross-modal classification. B-mode and Doppler modalities are combined using the cross-attention model and uncertainty-aware risk stratification, resulting in 96.4% accuracy, 94.6% sensitivity, 97.7% specificity, and 0.966 AUC on 1,990 images. CAFNet did better than single-modality baselines (+8.0 vs. B-mode, +9.7 vs. Doppler) and simple fusion (+5.2 vs. concatenation). The attention module acquired any meaningful morphological-hemodynamic agreement with high spatial agreement (Dice = 0.73) to expert annotations. The quantification of uncertainty was used to recognize 78 percent of misclassifications as high-uncertainty examples, which could be reviewed by experts, whereas low-risk predictions (72.5%) had 98.2% accuracy, which decreased the workload of radiologists by the same percentage. CAFNet can be deployed to the conventional ultrasound systems with real-time inference (23.8 FPS, 42 ms latency) and a 7.1 MB quantized model. Statistical significance proved a significantly greater superiority than TransUNet ($p = 0.008$) and clinical baselines ($p < 0.001$), which implies the possibility that it is an effective, clinically viable decision support tool that can be applied to ultrasound-based diagnosis.

Future Work: CAFNet creates a foundation of broader clinical adoption. The future work can focus on the multi-center validation between institutions, vendors, and populations in order to bring generalization and interoperability. The planned extensions are multi-class tumor classification, the combination with elastography and contrast-enhanced ultrasound, and the adaptation to other organs like the breast, thyroid, and liver. Interpretability and trust will be improved by a clinical interface that is interactive with attention maps and uncertainty visualization. CAFNet will also be assessed as a treatment monitoring and remote diagnostics tool in longitudinal and telemedicine studies.

Acknowledgments: The authors have no acknowledgments to declare.

Author Contributions: All authors contributed substantially and equally to the research conception, methodology development, data analysis, and manuscript preparation. All authors reviewed and approved the final version of the paper.

Data Availability: The datasets used in this study are publicly available: Uterine Fibroid Ultrasound Images dataset (Yang et al., 2023, Mendeley Data V2; DOI: 10.17632/n2zcmcypgb.2)

Ethics Statement: This study used publicly available, de-identified datasets (DOI: 10.17632/n2zcmcypgb.2). No institutional review board approval was required for secondary analysis of anonymized data per institutional policy.

Funding: The author(s) received no specific funding for this work.

Competing Interests: The author(s) have declared that no competing interests exist.

References

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of big Data, 8(1), 53. https://doi.org/10.1186/s40537-021-00444-8

Chen, H., Yang, B. W., Qian, L., Meng, Y. S., Bai, X. H., Hong, X. W., ... & Feng, W. W. (2022). Deep learning prediction of ovarian malignancy at US compared with O-RADS and expert assessment. Radiology, 304(1), 106-113. https://doi.org/10.1148/radiol.211367

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306. https://doi.org/10.48550/arXiv.2102.04306

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3146-3154). https://doi.org/10.1109/CVPR.2019.00326

Gal, Y., Islam, R., & Ghahramani, Z. (2017, July).

Deep bayesian active learning with image data. In International conference on machine learning (pp. 1183-1192). PMLR. https://doi.org/10.48550/arXiv.1703.02910

Ghosal, S., Xie, A., & Shah, P. (2021). Uncertainty quantified deep learning for predicting dice coefficient of digital histopathology image segmentation. arXiv preprint arXiv:2109.00115.

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., & Li, M. (2019). Bag of tricks for image classification with convolutional neural networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 558-567). https://doi.org/10.1109/CVPR.2019.00065

Huo, T., Li, L., Chen, X., Wang, Z., Zhang, X., Liu, S., ... & Deng, K. (2023). Artificial intelligence-aided method to detect uterine fibroids in ultrasound images: a retrospective study. Scientific Reports, 13(1), 3714. https://doi.org/10.1038/s41598-022-26771-1

Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P., & Kapur, T. (2020). Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE transactions on medical imaging, 39(12), 3868-3878. https://doi.org/10.1109/TMI.2020.3006437

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... & Rueckert, D. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999. https://doi.org/10.48550/arXiv.1804.03999

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., & Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis, 53, 197-207. https://doi.org/10.1016/j.media.2019.01.012

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of big data, 6(1), 1-48. https://doi.org/10.1186/s40537-019-0197-0

Song, X., Chao, H., Xu, X., Guo, H., Xu, S., Turkbey, B., ... & Yan, P. (2022). Cross-modal attention for multi-modal image registration. Medical Image Analysis, 82, 102612. https://doi.org/10.1016/j.media.2022.102612

Tan, M., & Le, Q. (2021, July). Efficientnetv2: Smaller models and faster training. In International conference on machine learning (pp. 10096-10106). PMLR. https://proceedings.mlr.press/v139/tan21a.html

Tao, S. Q., Shi, L. Y., Chai, X. S., Yuan, X. R., Tang, S. X., Zhang, J., ... & Fu, C. (2025). Value of multi-modal MRI in predicting the effect of high-intensity focused ultrasound for uterine fibroids. International Journal of Hyperthermia, 42(1), 2495360. https://doi.org/10.1080/02656736.2025.2495360

Toyohara, Y., Sone, K., Noda, K., Yoshida, K., Kurokawa, R., Tanishima, T., ... & Osuga, Y. (2022). Development of a deep learning method for improving diagnostic accuracy for uterine sarcoma cases. Scientific Reports, 12(1), 19612. https://doi.org/10.1038/s41598-022-23064-5

Volinsky-Fremond, S., Horeweg, N., Andani, S., Barkey Wolf, J., Lafarge, M. W., de Kroon, C. D., ... & Bosse, T. (2024). Prediction of recurrence risk in endometrial cancer with multimodal deep learning. Nature medicine, 30(7), 1962-1973. https://doi.org/10.1038/s41591-024-02993-w

Wang, K., Yin, Q., Wang, W., Wu, S., & Wang, L. (2016). A comprehensive survey on cross-modal retrieval. arXiv preprint arXiv:1607.06215. https://doi.org/10.48550/arXiv.1607.06215

Wang, X., Hu, Q., Dai, X., Li, P., & Chen, Y. (2023, November). Prediction of endometrial cancer MRI images based on deep learning and improved Bayesian extreme learning machine ensemble learning. In 2023 5th International Conference on Frontiers Technology of Information and Computer (ICFTIC) (pp. 560-565). IEEE. https://doi.org/10.1109/ICFTIC59930.2023.10456166

Xie, A., Elfatimi, E., Ghosal, S., & Shah, P. (2024, December). Deep learning with uncertainty quantification for predicting the segmentation dice coefficient of prostate cancer biopsy images. In 2024 International Conference on Machine Learning and Applications (ICMLA) (pp. 1158-1163). IEEE. https://doi.org/10.48550/arXiv.2109.00115

Yang, T., Yuan, L., Li, P., & Liu, P. (2023). Real-time automatic assisted detection of uterine fibroid in ultrasound images using a deep learning detector. Ultrasound in Medicine & Biology, 49(7), 1616-1626.

https://doi.org/10.17632/n2zcmcypgb.2
Zhao, F., Wang, Z., Du, H., He, X., & Cao, X. (2023). Self-supervised triplet contrastive learning for classifying endometrial histopathological images. IEEE Journal of Biomedical and Health Informatics, 27(12), 5970-5981. https://doi.org/10.1109/JBHI.2023.3314663
Zhou, T., Ruan, S., & Canu, S. (2019). A review: Deep learning for medical image segmentation using multi-modality fusion. Array, 3, 100004. https://doi.org/10.1016/j.array.2019.100004