



OPEN Deep learning-based classification of thyroid nodules using uncertainty-aware multi-modal ultrasound imaging

Manali Saini¹, Tanin Adl Parvar¹, Masiel Velarde¹, Nicholas B. Larson², Mostafa Fatemi³ & Azra Alizad^{1,3}✉

Accurate differentiation of thyroid nodules is crucial for timely diagnosis of thyroid cancer. Most recent studies utilize grayscale ultrasound with deep learning to distinguish benign from malignant thyroid nodules. The goal of this study is to effectively boost the performance of classification of thyroid nodules using multi-modal ultrasound imaging combining B-mode, color Doppler (CD), and shear wave elastography (SWE) within a customized deep learning architecture. This study prospectively included 506 thyroid nodules acquired from 422 subjects. The proposed network integrated a pretrained MobileNetV2 backbone with a shallow head composed of depth-wise separable convolutional layers, attention-based mixed pooling, and a tailored self-attention mechanism—applied for the first time in the context of thyroid nodule classification. To further improve robustness, we introduced a patient-level uncertainty-aware fusion strategy that selectively integrates predictions from each modality based on their validation scheme and achieved a classification accuracy of 0.95, a sensitivity (Sen) of 0.98, an area under the ROC curve (AUC) of 0.97 (95% CI: 0.94–0.99), a specificity of 0.92, and an F1 score of 0.95, on test data. Multi-modal images improved the performance (AUC: 0.97) as compared to uni-modal (AUC range: 0.73–0.90) or bi-modal (AUC range: 0.90–0.97) data. Further comparative analysis showed that the proposed network performed similarly or better to the state-of-art deep learning networks (AUC range: 0.82–0.97) for thyroid nodule classification while utilizing relatively smaller architecture. Finally, the integration of multi-modal ultrasound imaging with a customized deep learning network effectively and efficiently improved the overall classification of thyroid nodules, which can further enhance diagnostic performance.

Keywords Thyroid nodule classification, Deep learning, Ultrasound, Thyroid cancer, Shear wave elastography, Color doppler

Thyroid nodules are common findings in adults undergoing ultrasonography (US), with detection rates ranging from 10% to 67%¹. The majority of these nodules are benign, as fine needle aspiration biopsy (FNAB)—the gold standard for detecting thyroid malignancy—detects cancer in approximately 9.2% to 14.8% of cases². Given that thyroid cancer is the ninth most common cancer globally³ with approximately 44,000 new cases diagnosed annually in the United States, accurate differentiation of benign and malignant nodules is essential⁴.

Ultrasound is the first-line imaging modality for evaluating thyroid nodules due to its non-invasive nature, accessibility, and ability to provide detailed anatomical information^{5,6}. However, US imaging is highly operator-dependent and interpretation is subjective, which can lead to variability in diagnostic outcome⁷. These inconsistencies in diagnosis often result in unnecessary FNABs, which are invasive and cause patient discomfort and increase healthcare costs, highlighting the need for more reliable diagnostic methods^{8,9}.

Recent advances in multimodal ultrasound techniques, including B-mode, shear wave elastography (SWE), and color doppler (CD) have shown improvements in diagnostic accuracy. B-mode US provides details of morphological features of thyroid nodules; SWE evaluates tissue stiffness and color doppler assesses blood

¹Department of Radiology, Mayo Clinic College of Medicine and Science, 200 1st Street SW, Rochester, MN 55905, USA. ²Department of Quantitative Health Sciences, Mayo Clinic College of Medicine and Science, 200 1st St. SW, Rochester, MN 55905, USA. ³Department of Physiology & Biomedical Engineering, Mayo Clinic College of Medicine and Science, 200 1st St. SW 55905, Rochester, MN, USA. ✉email: Alizad.azra@mayo.edu

flow patterns^{10–14}. However, despite these advances, studies show that inconsistencies and controversies remain regarding the additional diagnostic value of multimodal ultrasound.

Artificial intelligence (AI)-powered machine learning (ML) has shown promising potential in addressing these limitations. AI can help reduce misdiagnoses caused by subjective biases of physicians, optimize task allocation, reduce workload, and improve the diagnostic accuracy of thyroid nodule assessments¹³. Previous studies have shown that ML¹⁵ and deep learning (DL) algorithms have achieved high accuracy in segmenting and classifying thyroid nodules as compared to experienced radiologists. For example, Chi et al. proposed fine-tuned GoogLeNet for extracting deep features from B-mode images of thyroid nodules and classifying them using a cost-sensitive random forest classifier¹⁶. Yadav et al. utilized computer aided diagnosis (CAD)-based convolutional and hybrid architectures to classify benign and malignant thyroid nodules from B-mode ultrasound in¹⁷. Prior to classification, they employed edge preserving smoothing de-speckling filter and encoder decoder-based ResNet50 for segmentation as pre-processing stages to enhance the performance of thyroid tumor characterization¹⁷. Wang et al. utilized a VGG16-based fine-tuned network to classify thyroid nodules using B-mode US images and compared the results with a radiomics-based approach¹⁸. Kim et al. compared the performance of three DL networks: VGG19, VGG16, ResNet with radiologists in discriminating thyroid nodules based on B-mode images¹⁹. Yang et al. also utilized ResNet18²⁰ for classifying B-mode images of thyroid nodules. Guan et al. explored Inception-v3 network for thyroid nodule classification using B-mode images in²¹. Some recent research studies utilized multi-modal ultrasound to improve thyroid nodule classification. For example²² et al. extracted radiomics features from SWE and B-mode images to classify thyroid nodules using ML classifiers²³. Liu et al. utilized the B-mode images with VGG16-based network and radiofrequency signals with EEGNet for thyroid nodule diagnosis after fusing the extracted features from both modalities²⁴. Recently, strain elastography and B-mode images were utilized in²⁵ with EfficientNet-B4 network for malignancy prediction of TI-RADS 4 thyroid nodules. Tao et al. integrated B-mode images, color Doppler flow imaging, strain elastography, and region of interest mask images to train independent ResNet-50 backbone networks with attention-based-fusion before the final layer for diagnosing suspicious thyroid nodules²⁶.

The aforementioned studies utilized multimodal US for thyroid nodule classification using deep learning. However, these studies are limited by either the use of less quantitative modalities such as strain elastography, restricted modality combinations, or the lack of uncertainty modeling, which is critical for trustworthy clinical deployment. To address these limitations, we propose a custom deep learning network that integrates B-mode, CD, and SWE modalities for thyroid nodule classification. Unlike previous approaches for thyroid nodule classification, our method employs a patient-level uncertainty-aware fusion ensemble to enhance diagnostic robustness and reliability. By explicitly modeling uncertainty in multi-modal predictions at patient-level, our approach aims to improve confidence estimation and support clinical decision-making, especially in borderline or ambiguous cases such as conflicting modalities. The aim of our study is to enhance diagnostic performance of US and ultimately improve patient care by integrating AI with multimodal ultrasound for thyroid nodule classification. Although the uncertainty-aware fusion of multiple modalities with deep learning has been employed recently in^{27,28} for eye disease screening and Covid-19 X-ray images classification respectively, this is the first work which proposes patient-level uncertainty-aware fusion of tri-modal (B-mode, SWE, CD) US with a custom deep learning architecture for thyroid nodule classification.

Methods

This prospective study was performed in accordance with the Declaration of Helsinki and all relevant guidelines and regulations under an approved Mayo Clinic Institutional Review Board (IRB), protocol (IRB: 08-008778), and in compliance with the Health Insurance Portability and Accountability Act (HIPAA). Written Mayo Clinic IRB-approved informed consent with permission for publication was obtained from all participants prior to the imaging study.

Study cohort

The study enrollment was conducted during 2016–2019 on consecutive patients with thyroid nodules identified in clinical ultrasound and recommended for FNAB. Patients aged eight years or above with one or more nodules were included in the study, whereas individuals with any history of thyroid surgery were excluded. Figure 1 summarizes the study population. A total of 31 cases with indeterminate pathology were excluded, leaving 422 patients for analysis. The mean (SD) age of 55.33 (14.22) years and the cohort included 122/422 (29%) male and 300/422 (71%), female. Thyroid Imaging Reporting and Data System (TI-RADS) scores were assigned by radiologists based on the clinical ultrasound features of nodules. Since several patients had multiple nodules, the total number of nodules evaluated was 506. Among the 422 patients, 64.7% had benign nodules, while 35.3% had malignant nodules. The benign nodules were labeled based on the results of FNAB and surgical pathology, while the final diagnoses for malignant nodules were based on surgical pathology.

Multimodal ultrasound imaging

For each nodule, two B-mode and six SWE images were acquired in each of the longitudinal and transverse orientations using a research-mode GE LOGIQ E9, (GE Healthcare, Wauwatosa, WI), equipped with a 9 L-D linear array transducer with frequency range of 2–8 MHz. The participants were instructed to restrict any body movement and pause respiration during each SWE acquisition for approximately 3 s, to minimize the pre-compression during scanning. Each SWE acquisition consisted of two simultaneously captured images placed side by side: a conventional B-mode image (left) and a SWE map (right), showing tissue stiffness values as a color-coded overlay superimposed on the corresponding B-mode image. This dual-view format allowed for both morphological assessment and elasticity evaluation in a single frame, facilitating comprehensive input for deep learning-based classification. The CD images were acquired using the clinical GE LOGIQ E9 machine, which

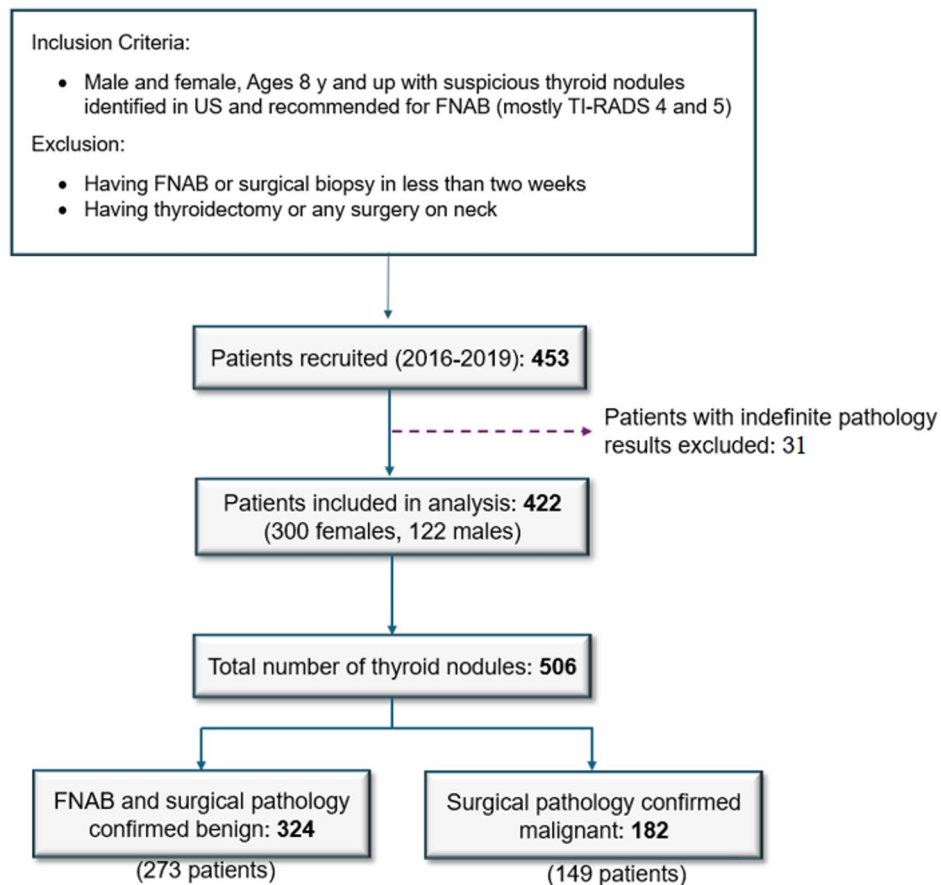


Fig. 1. Flowchart for the participants enrolled in this study.

operates with distinct acquisition settings and processing pipelines compared with the research platform. Up to two acquisitions per orientation were acquired. Each patient had varying number of images depending upon the nodules. A total of 6755 B-mode and SWE (combined) and 1406 CD images were obtained from all patients, reflecting multimodal imaging captured under two different acquisition environments within the GE LOGIQ E9 family. As an example, Fig. 2 shows the concatenated B-mode, SWE, and CD images acquired from two patients with benign and malignant nodules. It can be observed that the stiffness is higher for malignant nodules based on shear wave maps.

Proposed deep learning network

Each ultrasound image modality, i.e., B+SWE and CD was given to a separate instance of the proposed DL network, independently trained for each modality, as shown in Fig. 3. Since the number of images per patient was different, the images were grouped and fed into the model patient-wise, ensuring that all available images and nodules corresponding to a single patient were used collectively during training or testing. The proposed DL network comprised a pretrained MobileNetV2 architecture which extracted the high-level features from these images based on the learned weights from the ImageNet dataset. The corresponding learned features were fed to a customized shallow network (CSN) fine-tuned on our dataset for feature extraction and classification. The CSN architecture comprised majorly two blocks of two-dimensional depth-wise separable convolutional (DSC) and mixed pooling with attention (MPA) layers, followed by two DSC layers, a mixed pooling (MP) layer, tailored self-attention (TSA), flatten and a fully connected output layer with sigmoid activation. The first custom layer (i.e., DSC) utilized a depth-wise separable convolution operation with 128 filters, each with a size of 3×3 , and rectified linear unit (RELU) as the activation function. This is a two-step operation: in the first step, each channel is convolved with a filter at a time by splitting the filter into three 3×1 filters, and in the second step, standard convolution is performed with 1×1 filters.

The corresponding output feature map $f1_{DSC}$ of this layer can be represented as:

$$f1_{DSC} = DSC_{3 \times 3} (I_{MUS}) \quad (1)$$

where $DSC_{3 \times 3}$ denotes the depth-wise separable convolution operation with 3×3 filters, and (I_{MUS}) denotes the multi-modal US image input. Such 128 feature maps were obtained after this layer and further down-sampled by using mixed-pooling with attention (MPA) layer which used weighted sum of max-pooling and average pooling layer outputs based on sigmoid activation, as shown in Figure (3). This helped in generating

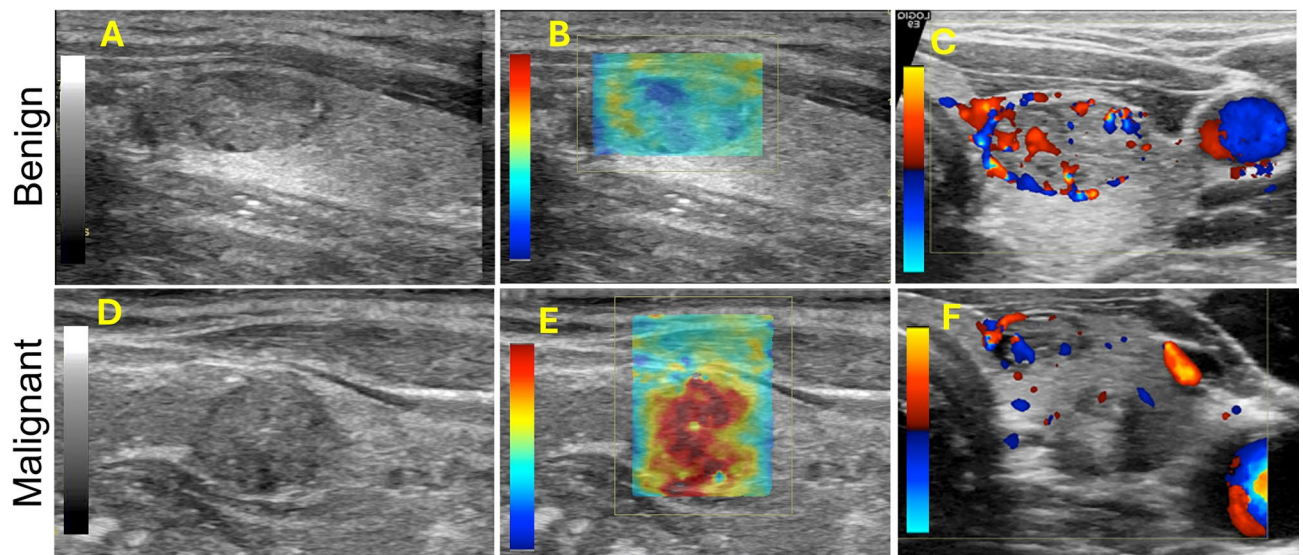


Fig. 2. Multimodal US images of two participants with Benign nodule (top, A: B-mode, B: SWE, and C: CD) and Malignant nodule (bottom, D: B-mode, E: SWE, and F: CD).

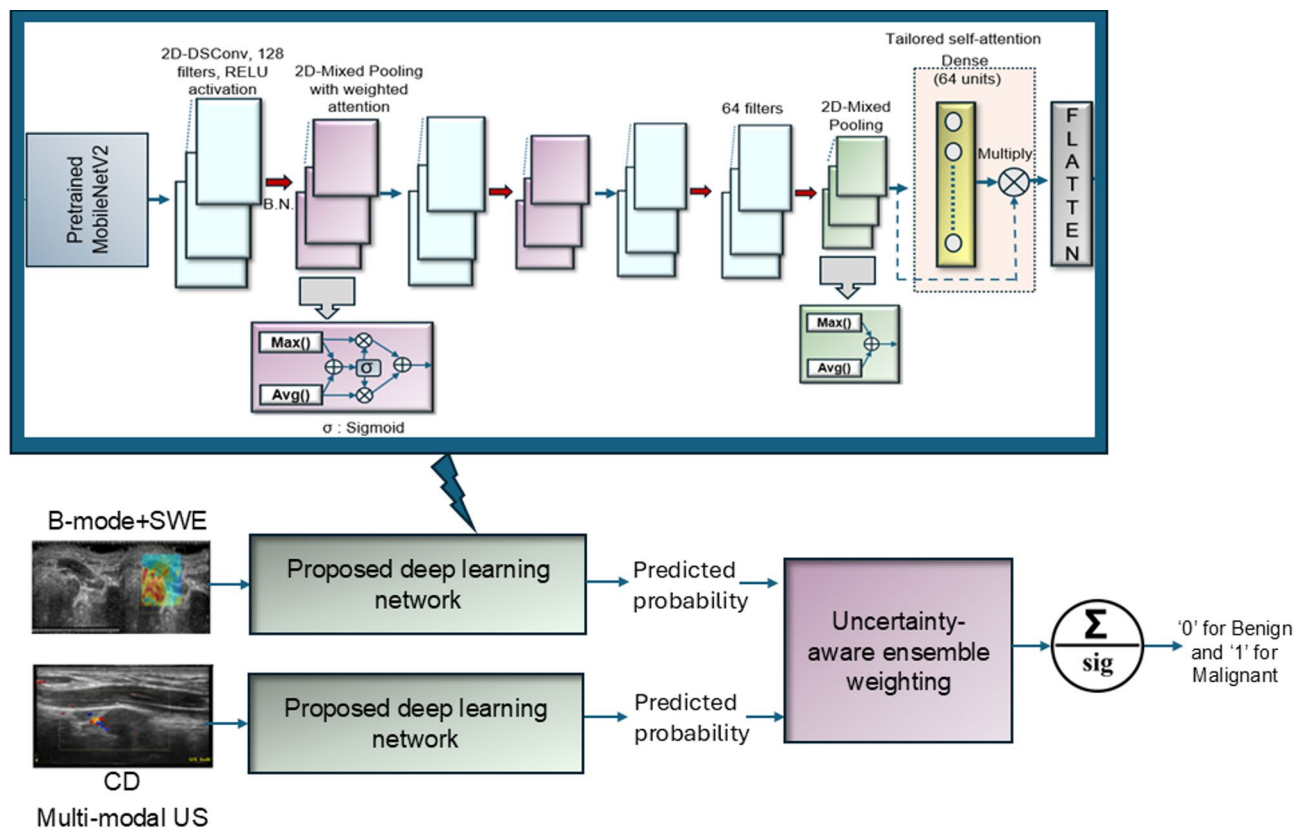


Fig. 3. Proposed deep learning architecture for categorizing thyroid nodules using multi-modal ultrasound. B.N. denotes batch normalization.

an enhanced feature map by combining relevant amounts of local and global information captured by max-pooling and average pooling layers respectively. This feature map can be represented as $f1_{MPA}$ and obtained as:

$$f1_{MPA} = \alpha (max_{2 \times 2} (f1_{DSC})) + (1 - \alpha) (avg_{2 \times 2} (f1_{DSC})) \quad (2)$$

where $max_{2 \times 2}$ and $avg_{2 \times 2}$ denote the max-pooling and average pooling operations with stride of 2×2 , and α is obtained as the sigmoid activated output of simple mixed pooling layer. As shown in Figure (3), there was a second block of DSC and MPA layers. Therefore, the feature map $f2_{MPA}$ was obtained after the second block in a similar manner as (1) and (2). This was given as input to two successive DSC layers with 128 and 64 filters each with a size of 3×3 , which produced the output feature maps $f3_{DSC}$ and $f4_{DSC}$ respectively. The feature map $f4_{DSC}$ was down sampled by using an MP layer (without attention). This helped in obtaining $f1_{MP}$, which had local and global information at the same level, as follows.

$$f1_{MP} = max_{2 \times 2} (f4_{DSC}) + avg_{2 \times 2} (f4_{DSC}) \quad (3)$$

After these operations, we introduced a tailored self-attention mechanism to further enhance the feature map for relevant areas in the image. TSA consisted of a dense/fully connected layer with 64 units, *RELU* activation, and a multiply layer. The preceding feature maps ($f1_{MP}$) were fed to the dense layer which produced the learned weights for each of the 64 maps by using element-wise multiplication²⁹. The corresponding weighted outputs were then multiplied with those initial 64 maps, thus producing a self-attention mechanism tailored to the input maps. Finally, all attention maps were flattened to a vector which was fed to a dense layer with sigmoid activation for classification.

The proposed architecture was built in TensorFlow Keras 2.16.1 deep learning framework in Python 3.10 Spyder environment. We augmented the dataset using random cropping, random rotation, and zooming techniques for including sufficient examples in training. The data was split into 80-10-10% for training-validating-testing. Training hyperparameters were selected using Keras Tuner and finalized based on performance on the validation set. The model was trained using the Adam optimizer with a learning rate of 0.0001. A batch size of 8 and a maximum of 50 epochs were used, with early stopping applied to prevent overfitting. A dropout rate of 0.3 was applied after the flatten layer. Binary cross-entropy was used as the loss function²⁹.

Each instance of the proposed network predicted the probability based on combined B-mode and SWE, and CD images. To enhance the robustness of multimodal fusion, we proposed an uncertainty-aware fusion framework that dynamically combined modality-specific predictions from each patient using confidence-informed decision rules. Instead of simply averaging or weighing the predictions, our method utilized the predictive uncertainty of each modality to guide the fusion process, allowing the model to favor more reliable predictions while down-weighting uncertain ones. We describe the proposed fusion process as follows.

Per-modality prediction and uncertainty estimation

We employed Monte Carlo (MC) Dropout³⁰ to estimate epistemic uncertainty in each modality-specific model instance for each patient. During inference, dropout layers were activated, and the model performed multiple stochastic forward passes. This yielded a distribution of predicted probabilities for both inputs, from which we computed the mean prediction ($\bar{p}$), as the final class probability and predictive variance (σ^2) as the uncertainty estimate. For a given input modality M , the model outputs prediction probabilities $\{p_1, p_2, \dots, p_N\}$ across N stochastic passes. The uncertainty is then estimated as the variance:

$$Uncertainty(M) = \sum_{\{n=1\}}^{\{N\}} (p_n - \bar{p})^2 \quad (4)$$

This uncertainty estimate highlights the network's confidence in its prediction.

Fusion strategy

In this work, each modality's contribution to the final prediction was weighted based on its predictive uncertainty, i.e., modalities with uncertainty below a predefined threshold (0.5) were given higher influence, while those exceeding this threshold were downweighted. The resulting fused probability was then compared against a separate classification threshold (set to 0.5) to determine the final class label. We used a classification threshold of 0.5 because the model outputs represent the predicted probability of malignancy after sigmoid activation, and 0.5 is the standard (default) decision boundary for binary classification under equal class weighting³¹. This adaptive fusion approach enhanced robustness to noisy or degraded modalities and contributes to more reliable clinical decision-making. The corresponding mathematical formulation is detailed as follows.

After the uncertainty estimation for each modality, the modality weight is defined as:

$$w(M) = \begin{cases} 1 - Uncertainty(M), & \text{if } U(M) \leq 0.5, \\ 0 & \text{if } Uncertainty(M) > 0.5 \end{cases}$$

The fused probability is then computed as a normalized weighted sum:

$$p_{fused} = \frac{(\sum_M w(M) \cdot \bar{p}^M)}{(\sum_M w(M))}$$

Modality	Acc	Sen	Spe	F 1	AUC
B-mode	0.85 (0.09)	0.79 (0.10)	0.89 (0.08)	0.80 (0.12)	0.89 (0.79–0.92)
CD	0.73 (0.16)	0.51 (0.13)	0.84 (0.13)	0.56 (0.11)	0.73 (0.66–0.78)
SWE	0.87 (0.09)	0.91 (0.10)	0.90 (0.09)	0.88 (0.11)	0.90 (0.81–0.93)
B-mode + CD	0.86 (0.09)	0.81 (0.08)	0.90 (0.11)	0.81 (0.12)	0.90 (0.83–0.93)
CD + SWE	0.88 (0.10)	0.92 (0.12)	0.91 (0.11)	0.89 (0.13)	0.91 (0.81–0.93)
B-mode + SWE	0.93 (0.10)	0.91 (0.08)	0.91 (0.08)	0.91 (0.11)	0.97 (0.92–0.99)
Combined	0.95 (0.09)	0.98 (0.07)	0.92 (0.10)	0.95 (0.09)	0.97 (0.94–0.99)

Table 1. Classification performance metrics across different folds (Mean (S.D.) and AUC (95% C.I.) of the proposed network for uni-modal and multi-modal data.

Fusion strategy	Acc	Sen	Spe	F1	AUC
Weighted Average	0.92 (0.10)	0.95 (0.09)	0.86 (0.08)	0.91 (0.09)	0.94 (0.88–0.96)
Max confidence	0.82 (0.12)	0.85 (0.11)	0.81 (0.11)	0.80 (0.06)	0.84 (0.79–0.87)
Uncertainty-Aware (Inverse)	0.81 (0.10)	0.75 (0.09)	0.87 (0.10)	0.82 (0.08)	0.84 (0.78–0.88)
Uncertainty-Aware (Proposed)	0.95 (0.09)	0.98 (0.07)	0.92 (0.10)	0.95 (0.09)	0.97 (0.94–0.99)

Table 2. Classification performance metrics across different folds (Mean (S.D.)) and AUC (95% CI) of the proposed network with different ensemble fusion techniques. Values that equal or exceed the proposed network are bolded.

This explicit formulation clarifies how uncertainty directly modulates each modality’s contribution to the final classification.

Model training and performance evaluation

A five-fold nested cross-validation design was used for model training and testing, and all assessments were based on patient-level predictions. For each outer-fold, the complementary 80% of patients were used for model training and the trained network was used to classify the leave-out test set consisting of 20% of the patients in each fold. Classification performance was evaluated in terms of Accuracy (Acc), sensitivity/recall (Sen), specificity (Spe), and F1 score, which were summarized based on means and standard deviations (SD) of the classification metrics obtained across the five folds. Discrimination was evaluated based on area under the receiver operating characteristic curve (AUC), with cross-validated AUC and 95% confidence intervals (CIs) estimated using the cvAUC method³².

Results

Performance assessments for models trained with all possible combinations of image modalities, including single and bi-modal images, are summarized in Table 1. Among the single modalities, SWE numerically demonstrated the highest performance across all metrics, although AUC 95% CIs for B-mode and SWE heavily overlapped. Amongst dual modalities, the combination of B-mode and SWE consistently demonstrated higher performance metric point estimates, specifically in terms of Acc of 0.93 and AUC of 0.97 (95% CI: 0.92–0.99). Finally, the proposed architecture for all three modalities achieved the best performance, with the AUC 95% CI demonstrating no overlap with any of the uni-modal networks.

Performance analysis with respect to different ensemble fusion techniques

Next, we evaluated weighted averaging, maximum confidence, and inverse uncertainty techniques³³. Table 2 shows comparative performance metrics which demonstrate that the proposed uncertainty-aware threshold-based fusion higher evaluation metrics, achieving an accuracy of 0.95 and an AUC of 0.97. We observed non-overlapping CIs of the max confidence and inverse uncertainty aware techniques relative to the proposed fusion strategy. This indicates its effectiveness in utilizing multi-modal data while suppressing the impact of uncertain or conflicting modality predictions. The performance was otherwise comparable to the weighted average fusion technique. These results provide direct validation of the uncertainty-aware mechanism. Because the fusion weight $w(M)$ decreases as predictive variance increases, modalities with unstable or noisy predictions had reduced influence on the fused decision. The superior AUC (0.97) and non-overlapping confidence intervals relative to the maximum-confidence and inverse-uncertainty baselines demonstrate that this uncertainty-driven modulation improved the performance.

Ablation study on network components

To demonstrate the effectiveness of the proposed network, we conducted a detailed ablation study by systematically removing or replacing its key components. The results, summarized in Fig. 4, highlight the performance impact of removing/replacing each key component across multiple evaluation metrics. The full proposed model achieved the highest performance, with an accuracy of 0.95 and an AUC of 0.97. When pretraining was removed, performance dropped notably, indicating the importance of utilizing pretrained weights for effective feature extraction. Similarly, replacing the tailored self-attention mechanism with a standard convolutional block resulted in a reduction in F1 score and sensitivity, indicating the role of attention in capturing discriminative modality-specific features. Eliminating mixed pooling also degraded the performance, indicating that the combination of max and average pooling improves generalization by preserving diverse feature information. Among pretrained backbone variants, EfficientNetB0 and NASNetMobile resulted in lower performance compared to the proposed MobileNetV2 network, demonstrating the importance of selecting a suitable lightweight yet effective pretrained backbone for ultrasound data. These findings validate the architectural design choices and underscore the importance of each component in achieving robust multimodal classification performance.

Model interpretability via attention maps and uncertainty visualization

To evaluate the interpretability of the proposed network, we extracted tailored self-attention attention feature/heat-maps from each of B-mode + SWE and Color Doppler branches along with the modality-level uncertainty values used in the final uncertainty-weighted fusion. Figure 5 shows a benign thyroid nodule of a patient, with corresponding input images, attention heatmaps, estimated uncertainties and modality weights, and the final predicted classification probability (fused). B + SWE branch focuses on the nodule region and stiffness map, as indicated by the attention output, and outputs the correct prediction with a low uncertainty value (high confidence) of 0.12. However, the color Doppler shows somewhat scattered intensities in the attention map, and outputs false prediction (i.e., malignant) with a high uncertainty value (low confidence) of 0.84, which leads to zero weight as per the uncertainty-aware mechanism. Finally, the model outputs correct prediction based on only the highly confident B + SWE modality and final fused probability of 0.05 (benign).

These visualizations demonstrate that the proposed model effectively identifies nodule-specific regions via the tailored self-attention mechanism; unreliable modalities and down-weights or ignores them, relying on the most informative branches for robust final predictions. Importantly, even when CD predicts incorrectly, the uncertainty-aware ensemble prevents it from dominating the final classification, ensuring clinically reliable outputs.”

Performance comparison with state-of-the-art methods

We performed a comparative analysis of the proposed network with respect to some of the commonly used deep learning architectures for thyroid nodule classification. Since the datasets were different and mostly based on uni-modal images (B-mode US) in the reported studies, we implemented these common DL architectures

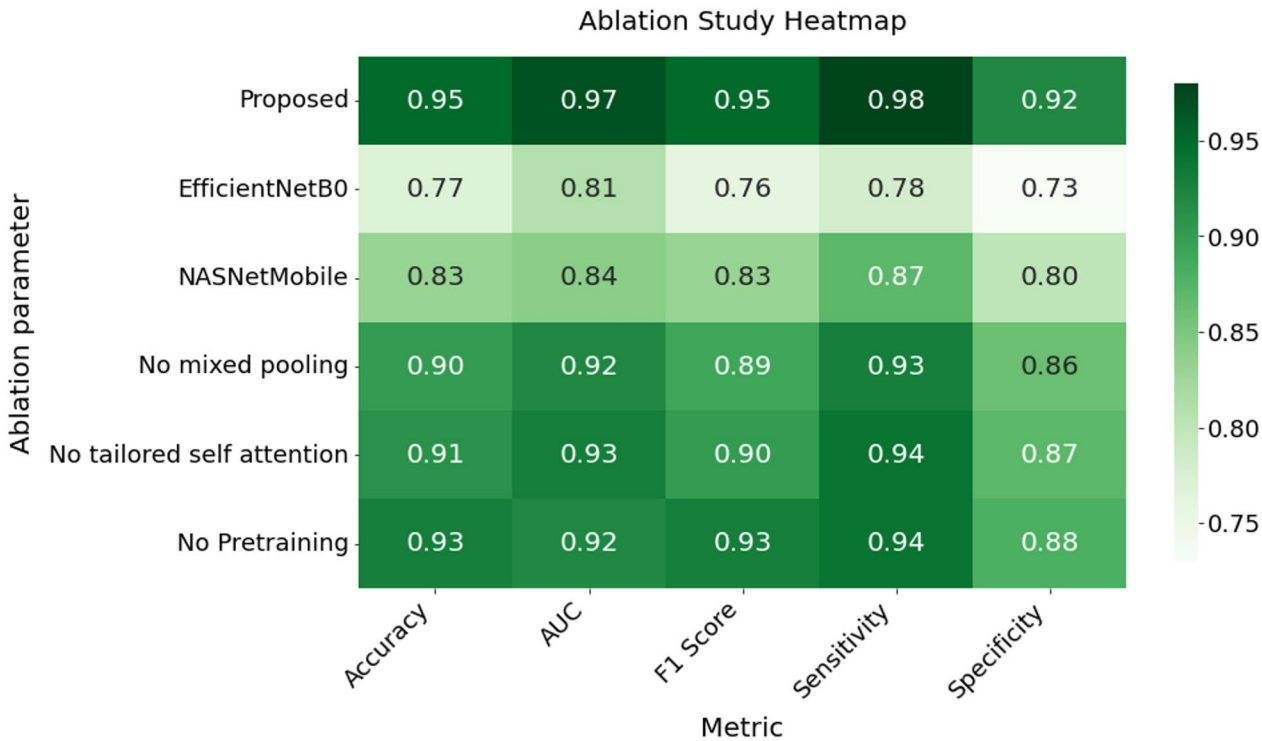


Fig. 4. Ablation analysis of the proposed network.

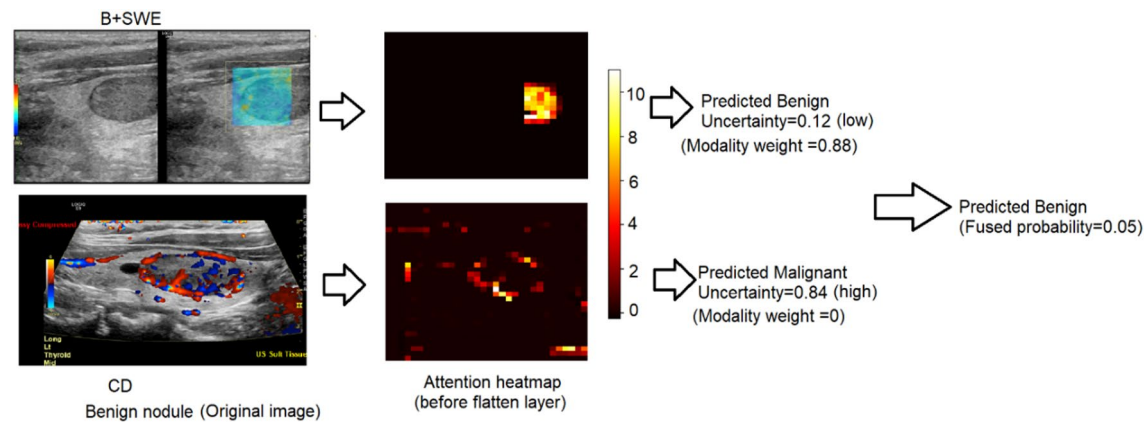


Fig. 5. An example of model interpretability. Left: Benign B-mode + SWE and CD input images. Middle: corresponding tailored self-attention maps. Right: predicted modality-level probability and uncertainty, assigned weight, and final fused prediction.

Method	Acc	Sen	Spe	F1	AUC
ResNet-50 with attention ²⁶	0.90	0.84	0.92	0.89	0.93 (0.84–0.95)
ResNet-34 ³⁴	0.82	0.85	0.81	0.80	0.84 (0.76–0.87)
YOLOv3 ³⁵	0.81	0.75	0.87	0.82	0.84 (0.78–0.88)
Fine-tuned GoogleNet ¹⁶	0.93	0.94	0.92	0.93	0.96 (0.87–0.98)
Fine-tuned VGG16 ^{18,19}	0.79	0.77	0.86	0.74	0.82 (0.76–0.85)
ResNet-18 ²⁰	0.94	0.93	0.95	0.94	0.97 (0.90–0.99)
Inception-v3 ²¹	0.93	0.91	0.94	0.93	0.93 (0.86–0.95)
DenseNet-121 ²²	0.87	0.89	0.88	0.85	0.89 (0.83–0.91)
DenseNet-169 ²²	0.81	0.83	0.80	0.80	0.83 (0.74–0.88)
NASNetLarge ²²	0.83	0.85	0.87	0.83	0.84 (0.77–0.89)
Exception ²²	0.90	0.86	0.90	0.88	0.91 (0.83–0.94)
ResNet101v2 ³⁶	0.86	0.87	0.81	0.86	0.88 (0.81–0.92)
ResNet152 ³⁶	0.91	0.89	0.86	0.90	0.93 (0.87–0.96)
MobileNetv2 ³⁶	0.91	0.90	0.87	0.91	0.94 (0.86–0.97)
Proposed	0.95	0.98	0.92	0.95	0.97 (0.94–0.99)

Table 3. Classification performance metrics across different folds (Mean) and AUC (95% CI) of the proposed network for multi-modal data. Values that equal or exceed the proposed network are bolded.

on our multi-modal US dataset, for fair comparison. We implemented and trained these state-of-the-art deep learning networks on the multi-modal data and evaluated them on the test set, using the same experimental settings and hyper-parameters of our proposed network. Further, all models were extended to tri-modal input using the same uncertainty-aware fusion mechanism. This ensured that performance differences reflect architectural design rather than fusion strategy. Table 3 presents the cross-validated classification metrics of these architectures along with the proposed network. The Acc values achieved by these architectures were in the range of 0.79–0.94 vs. 0.95 achieved by the proposed network, while the mean Sen values ranged between 0.75 and 0.93 vs. 0.98 respectively. Similarly, the values of F1 and AUC for existing networks were in the ranges of 0.74–0.94 and 0.82–0.97, respectively, vs. 0.95 and 0.97 for the proposed network respectively. We observed a similar trend in the values of specificities, except for two existing architectures, i.e., ResNet-18 and Inception-V3 achieving slightly higher values than the proposed network. We observed non-overlapping CIs for AUC for the existing networks ResNet-34, YOLOv3, Fine-tuned VGG-16, DenseNet-169, NASNetLarge, Exception, and ResNet101v2 relative to the proposed network, indicating its superiority in discrimination. Performance was otherwise comparable with respect to the other networks.

Discussion

In this study, we proposed multi-modal US-based customized deep learning architecture for thyroid nodule classification. For the first time in this context, we proposed the use of mixed pooling with weighted attention and tailored self-attention mechanisms on top of a pretrained backbone, MobileNetv2, in deep learning architecture. Additionally, we proposed patient-level uncertainty-aware fusion strategy to enhance the robustness of predictions obtained from each modality. An extensive performance assessment with respect to training the

proposed architecture instances using different US modalities and their combinations demonstrated that the uncertainty-aware fusion of all three modalities outperformed all other cases, in terms of all metrics. These findings suggest that integrating multi-modal US data considerably improves the discriminatory capability of the proposed architecture for thyroid nodule classification. While the combination of B-mode and SWE yielded competitive results, the inclusion of CD images further enhanced overall performance. Our findings also highlight the effectiveness of the proposed uncertainty-aware threshold-based fusion strategy, which consistently outperformed other ensemble techniques, demonstrating its robustness in handling uncertain or conflicting predictions across modalities. The ablation study confirmed the effectiveness of each component in the proposed architecture. These findings collectively validate the architectural design choices for robust multi-modal thyroid nodule classification. Specifically, a considerable rise in sensitivity and both AUCs depicted the precision of the proposed method in identifying most of the malignant nodules while achieving a minimal trade-off in the specificity, indicating its strong diagnostic performance.

The proposed method outperformed a recently proposed multi-modal US-based DL method utilizing ResNet-50 backbone with attention mechanisms²⁶ in terms of overall classification performance on our dataset, with a substantial increase in sensitivity, resulting in superior diagnostic performance. Further analysis of implementation of the state-of-the-art DL architectures for thyroid nodule classification on our multimodal dataset demonstrated superior performance, and fine-tuned GoogleNet, ResNet-18, and Inception-v3 achieved relatively close results with respect to proposed method^{16,20,21}. However, the proposed method demonstrated superior overall performance, while utilizing relatively few layers with effective attention mechanisms in its architecture, indicating its efficiency and applicability in real-time.

In a clinical workflow, uncertainty quantification provides radiologists with additional information beyond a binary model output. Rather than forcing a high-confidence label for every modality, the model highlights predictions associated with elevated epistemic uncertainty, which often corresponds to unusual appearances, degraded input quality, or conflicting information across modalities. These cases can be flagged for extra review, checked again by another radiologist, or lead to a request for more imaging. Conversely, low-uncertainty predictions provide radiologists with a more reliable initial assessment, potentially reducing diagnostic workload by prioritizing cases in routine workflows. In this way, the uncertainty-aware fusion used in this study directly supports practical clinical decision-making by improving interpretability, reducing the likelihood of overconfident errors, and enhancing trust in model outputs. The proposed uncertainty-aware mechanism therefore aligns with how radiologists naturally reason about diagnostic ambiguity, providing a quantified indication of model confidence that can inform safe and efficient clinical decisions. For example, if the ultrasound modality is noisy or incomplete, the elevated predictive variance results in its down weighting, preventing misleading modality-specific predictions from dominating the final fused output: mirroring the way radiologists give less weight to the unreliable views in practice.

In summary, this study demonstrates that multi-modal US imaging with proposed deep learning architecture can effectively classify thyroid nodules and boost diagnostic performance. By integrating uncertainty-aware fusion, the model not only achieves high accuracy but also improves reliability in the presence of conflicting or ambiguous modality inputs. This study needs to be further evaluated on an independent dataset. However, currently, no publicly available databases provide combined B-mode, SWE, and Color Doppler ultrasound for thyroid nodule assessment, which restricts its multi-center external validation. Therefore, our evaluation is limited to a single-institution dataset. Nevertheless, the uncertainty-aware fusion strategy implemented in our model is explicitly designed to mitigate the influence of unreliable or device-dependent modalities, which may enhance robustness when applied to data acquired from different ultrasound machines or clinical settings. Future work will focus on validating the proposed framework on multi-institutional datasets and across heterogeneous ultrasound devices to more comprehensively assess generalizability and clinical applicability. Further, we plan to acquire more images for increasing the number of examples during training to further improve performance. With sufficient data and continued refinement, the trained network can be deployed for real-time thyroid nodule classification, offering a reliable decision-support tool in clinical settings.

Conclusion

This study proposes multi-modal US-based deep learning architecture for the classification of thyroid nodules. The concatenation of conventional B-mode US with shear wave elastography and color Doppler images demonstrates a considerable enhancement in the performance of the proposed custom deep learning architecture with tailored attention mechanisms. Further comparative evaluation with state-of-art deep learning networks demonstrates the superiority and efficiency of the proposed network for thyroid nodule classification.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request. The requested data may include figures that have associated raw data. Because the study was conducted on human volunteers, the release of patient data may be restricted by Mayo policy and needs special requests. The request can be sent to: Karen A. Hartman, MSN, CHRC — Administrator - Research Compliance— Integrity and Compliance Office — Assistant Professor of Health Care Administration, Mayo Clinic College of Medicine & Science — 507-538- 5238 — Administrative Assistant: 507-266-6286 — hartman.karen@mayo.edu Mayo Clinic — 200 First Street SW — Rochester, MN 55905 — mayoclinic.org. We do not have publicly available accession codes, unique identifiers, or web links.

Received: 4 October 2025; Accepted: 9 January 2026

Published online: 12 January 2026

References

- Kim, J. Y., Jung, S. L., Kim, M. K., Kim, T. J. & Byun, J. Y. Differentiation of benign and malignant thyroid nodules based on the proportion of sponge-like areas on ultrasonography: imaging-pathologic correlation. *Ultrasonography* **34**, 304 (2015).
- Moon, W. J. et al. Benign and malignant thyroid nodules: US differentiation—multicenter retrospective study. *Radiology* **247**, 762–770. <https://doi.org/10.1148/radiol.2473070944> (2008).
- Chen, D. W., Lang, B. H. H., McLeod, D. S. A., Newbold, K. & Haymart, M. R. Thyroid cancer. *Lancet* **401**, 1531–1544. [https://doi.org/10.1016/S0140-6736\(23\)00020-X](https://doi.org/10.1016/S0140-6736(23)00020-X) (2023).
- Ginzberg, S. P. et al. Revisiting the relationship between tumor size and risk in Well-Differentiated thyroid cancer. *Thyroid* **34**, 980–989. <https://doi.org/10.1089/thy.2023.0327> (2024).
- Cibas, E. S. & Ali, S. Z. The 2017 Bethesda system for reporting thyroid cytopathology. *Thyroid* **27**, 1341–1346. <https://doi.org/10.1089/thy.2017.0500> (2017).
- Haugen, B. R. 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: What is new and what has changed? *Cancer* **123**, 372–381. <https://doi.org/10.1002/cncr.30360> (2017).
- Durante, C. et al. The diagnosis and management of thyroid nodules: A review. *JAMA* **319**, 914–924. <https://doi.org/10.1001/jama.2018.0898> (2018).
- Zhou, H. et al. Differential diagnosis of benign and malignant thyroid nodules using deep learning radiomics of thyroid ultrasound images. *Eur. J. Radiol.* **127**, 108992. <https://doi.org/10.1016/j.ejrad.2020.108992> (2020).
- Jeong, E. Y. et al. Computer-aided diagnosis system for thyroid nodules on ultrasonography: diagnostic performance and reproducibility based on the experience level of operators. *Eur. Radiol.* **29**, 1978–1985. <https://doi.org/10.1007/s00330-018-5772-9> (2019).
- Zhao, D. B., Jing, Y., Lin, X. Y. & Zhang, B. X. The value of color doppler ultrasound in the diagnosis of thyroid nodules: a systematic review and meta-analysis. *Gland Surg.* **10**, 3369–3377. <https://doi.org/10.21037/gs-21-752> (2021).
- Hang, J. et al. Combination of maximum shear wave elasticity modulus and TIRADS improves the diagnostic specificity in characterizing thyroid nodules: A retrospective study. *Int. J. Endocrinol.* **2018**, 4923050. <https://doi.org/10.1155/2018/4923050> (2018).
- Kim, H., Kim, J. A., Son, E. J. & Youk, J. H. Quantitative assessment of shear-wave ultrasound elastography in thyroid nodules: diagnostic performance for predicting malignancy. *Eur. Radiol.* **23**, 2532–2537. <https://doi.org/10.1007/s00330-013-2847-5> (2013).
- Kohlenberg, J. et al. Added value of mass characteristic frequency to 2-D shear wave elastography for differentiation of benign and malignant thyroid nodules. *Ultrasound. Med. Biol.* **48**, 1663–1671 (2022).
- Gregory, A. et al. Differentiation of benign and malignant thyroid nodules by using comb-push ultrasound shear elastography: a preliminary two-plane view study. *Acad. Radiol.* **25**, 1388–1397 (2018).
- Yadav, N., Dass, R. & Virmani, J. A systematic review of machine learning based thyroid tumor characterisation using ultrasonographic images. *J. Ultrasound.* **27**, 209–224 (2024).
- Chi, J. et al. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. *J. Digit. Imaging.* **30**, 477–486 (2017).
- Yadav, N., Dass, R. & Virmani, J. Deep learning-based CAD system design for thyroid tumor characterization using ultrasound images. *Multimedia Tools Appl.* **83**, 43071–43113 (2024).
- Wang, Y. et al. Comparison study of radiomics and deep learning-based methods for thyroid nodules classification using ultrasound images. *Ieee Access.* **8**, 52010–52017 (2020).
- Kim, Y. J. et al. Deep convolutional neural network for classification of thyroid nodules on ultrasound: comparison of the diagnostic performance with that of radiologists. *Eur. J. Radiol.* **152**, 110335 (2022).
- Yang, J. et al. Ultrasound image classification of thyroid nodules based on deep learning. *Front. Oncol.* **12**, 905955 (2022).
- Guan, Q. et al. Deep learning based classification of ultrasound images for thyroid nodules: a large scale of pilot study. *Annals Translational Med.* **7**, 137 (2019).
- Chen, C. et al. Deep learning approaches for differentiating thyroid nodules with calcification: a two-center study. *BMC Cancer.* **23**, 1139 (2023).
- Zhao, C. K. et al. A comparative analysis of two machine Learning-Based diagnostic patterns with thyroid imaging reporting and data system for thyroid nodules: diagnostic performance and unnecessary biopsy rate. *Thyroid* **31**, 470–481. <https://doi.org/10.1007/s00330-020-07585-z> (2021).
- Liu, Z. et al. Thyroid nodule recognition using a joint convolutional neural network with information fusion of ultrasound images and radiofrequency data. *Eur. Radiol.* **31**, 5001–5011. <https://doi.org/10.1007/s00330-020-07585-z> (2021).
- Chu, X. et al. Deep learning model for malignancy prediction of TI-RADS 4 thyroid nodules with high-risk characteristics using multimodal ultrasound: A multicentre study. *Comput. Med. Imaging Graphics* **102576** (2025).
- Tao, Y. et al. Deep learning for the diagnosis of suspicious thyroid nodules based on multimodal ultrasound images. *Front. Oncol.* **12**, 1012724 (2022).
- Zou, K. et al. Confidence-aware multi-modality learning for eye disease screening. *Med. Image. Anal.* **96**, 103214 (2024).
- Gour, M. & Jain, S. Uncertainty-aware convolutional neural network for COVID-19 X-ray images classification. *Comput. Biol. Med.* **140**, 105047 (2022).
- Saini, M., Afrin, H., Sotoudehnia, S., Fatemi, M. & Alizad, A. DMAeEDNet: dense multiplicative attention enhanced encoder decoder network for ultrasound-based automated breast lesion segmentation. *IEEE Access* (2024).
- Gal, Y. & Ghahramani, Z. in *international conference on machine learning*. 1050–1059 (PMLR).
- Goodfellow, I. (MIT Press, 2016).
- LeDell, E., Petersen, M. & van der Laan, M. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electron. J. Stat.* **9**, 1583 (2015).
- Gawlikowski, J. et al. A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.* **56**, 1513–1589 (2023).
- Qian, T. et al. Diagnostic value of deep learning of multimodal imaging of thyroid for TI-RADS category 3–5 classification. *Endocrine* **1–10** (2025).
- Zhang, X., Jia, C., Sun, M. & Ma, Z. The application value of deep learning-based nomograms in benign–malignant discrimination of TI-RADS category 4 thyroid nodules. *Sci. Rep.* **14**, 7878 (2024).
- Agyekum, E. A. et al. Ultrasound-based classification of follicular thyroid cancer using deep convolutional neural networks with transfer learning. *Sci. Rep.* **15**, 21708 (2025).

Acknowledgements

The authors would like to thank all past and present members, sonographers, and study coordinators who helped for a period during the years of this study.

Author contributions

Manali Saini wrote the original draft of manuscript, contributed to data curation, methodology, visualization, formal analysis, software, development of algorithms, reviewing and editing the manuscript; Tanin Adl Parvar

wrote the original draft of Introduction of the manuscript, contributed to data curation, visualization, reviewing and editing the manuscript; Masiel Velarde contributed to data curation, validation and reviewing editing the manuscript; Nicholas B. Larson contributed to formal data analysis, validation and reviewing and editing the manuscript; Mostafa Fatemi contributed to conceptualization, methodology, visualization, resources, funding acquisition, supervision, and project administration, reviewing and editing the manuscript; Azra Alizad contributed to conceptualization, methodology, investigation, visualization, resources, funding acquisition, supervision, project administration, reviewing and editing the manuscript. All authors reviewed the manuscript.

Funding statement

This work was supported in part by grants from the National Cancer Institute, R01CA239548 (Azra Alizad and Mostafa Fatemi), and the National Institute of Biomedical Imaging and Bioengineering, R01EB017213 (Azra Alizad and Mostafa Fatemi), both from the National Institutes of Health (NIH). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The NIH did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

Mayo Clinic Institutional Review Board approval (IRB: 08-008778) was obtained in compliance with the Health Insurance Portability and Accountability Act.

Informed consent

A signed written IRB approved informed consent was obtained from each enrolled participant prior to the prospective study.

Consent for publication

A signed written IRB approved informed consent was obtained with permission for publication.

Additional information

Correspondence and requests for materials should be addressed to A.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026