Title

Patient-level thyroid cancer classification using attention multiple instance learning on fused multi-scale ultrasound image features.

Authors

Zhuang, Luoting
Ivezic, Vedrana
Feng, Jeffrey
et al.

Publication Date

2023

Peer reviewed

Patient-level thyroid cancer classification using attention multiple instance learning on fused multi-scale ultrasound image features

Luoting Zhuang[1*], Vedrana Ivezic[1*], Jeffrey Feng[1*], Chushu Shen[2], Ashwath Radhachandran[2], Vivek Sant, MD[3], Maitraya Patel, MD[4], Rinat Masamed, MD[4], Corey Arnold, PhD[1,2,4], William Speier, PhD[1,2,4]

[1]Medical Informatics Home Area, University of California, Los Angeles, CA, USA;
[2]Department of Bioengineering, University of California, Los Angeles, CA, USA;
[3]Section of Endocrine Surgery, Department of Surgery, University of California, Los Angeles, CA, USA;
[4]Department of Radiological Sciences, University of California, Los Angeles, CA, USA

Abstract
For patients with thyroid nodules, the ability to detect and diagnose a malignant nodule is the key to creating an appropriate treatment plan. However, assessments of ultrasound images do not accurately represent malignancy, and often require a biopsy to confirm the diagnosis. Deep learning techniques can classify thyroid nodules from ultrasound images, but current methods depend on manually annotated nodule segmentations. Furthermore, the heterogeneity in the level of magnification across ultrasound images presents a significant obstacle to existing methods. We developed a multi-scale, attention-based multiple-instance learning model which fuses both global and local features of different ultrasound frames to achieve patient-level malignancy classification. Our model demonstrates improved performance with an AUROC of 0.785 ($p<0.05$) and AUPRC of 0.539, significantly surpassing the baseline model trained on clinical features with an AUROC of 0.667 and AUPRC of 0.444. Improved classification performance better triages the need for biopsy.

Introduction
Thyroid nodules are circumscribed solid or fluid-filled tissue of differential composition within the thyroid that may have malignant potential. If a suspicious neck mass is found or if symptoms of thyroid dysfunction are identified during a physical examination, patients typically undergo an ultrasound (US) exam for further investigation. US imaging has been widely used as a non-invasive, inexpensive, and real-time method for nodule detection. Palpable nodules are detected in 5%-7% of adults during a physical examination. On the other hand, the rate of nodule detection using US images is much higher and increases with age, with nodules detected in up to 70% of the population[1,2]. Discovered nodules are assessed and assigned a malignancy risk score ranging from 1 (benign) to 5 (highly suspicious) defined by the American College of Radiology Thyroid Imaging Reporting and Data Systems (ACR TI-RADS), which is calculated from the sum of points using nodule composition, echogenicity, shape, margin, and echogenic foci[3,4]. Each score has a corresponding recommendation based on the size of the nodule, with suspicious nodules recommended to undergo fine needle aspiration biopsy (FNAB) that samples cells for cytopathologists to analyze. Cytopathologic risk of malignancy follows the Bethesda System using the following categories: non-diagnostic (too few cells for diagnosis), benign, atypia of undetermined significance, suspicious for follicular neoplasm, suspicious for malignancy, and malignant[5,6]. A significant concern is that the majority of nodules undergoing FNAB are determined to be benign, suggesting that TI-RADS scores from the US images do not accurately indicate the malignancy risk determined after FNAB[7,8]. While FNAB is minimally invasive, small risks of bleeding and infections do exist, and for some patients, these procedures can be painful and extremely anxiety provoking. Compounded with the high rate of benign biopsies, a clear need arises for an improved tool to accurately predict the cytology of a nodule from the imaging data to better triage the need for biopsy.

The use of machine learning to classify cytology from thyroid US images has grown in popularity in recent years, and existing work has explored the application of both traditional machine learning and deep learning methods[9,10]. Many previous approaches include segmenting nodules from the US images, extracting radiomic features from the segmentations, and classifying the nodule as either benign or malignant[11,12]. A major limitation of such an approach is that it requires a large amount of segmented data for model training. Obtaining such data requires manual annotations from radiologists, which is a time-consuming and labor-intensive procedure. Consequently, there is

usually limited segmented data, which influences the learning and generalization of these models. Even current approaches that leverage weak supervision to achieve thyroid nodule classification depend on radiologist-annotated bounding boxes around nodules[13,14]. To circumvent this limitation, images without segmentations have been used as input into classification models[15,16]. An additional limitation of such approaches results from the use of US images which generally display information at different sizes and scales of magnification; yet, deep learning methods generally require consistent input formats. As a result, distortions due to image pre-processing can impact the portrayal of pertinent image information, and impact model training as well as generalizability.

The goal of this study was to build a weakly supervised classification model based on multiple-instance learning (MIL)[17,18] to predict the nodule cytology at the patient-level using fused information from all frames of a given patients' US scan. MIL has been used successfully for other problems such as breast cancer classification[19], Barett's cancer detection[20], and lung cancer diagnosis[21]. The benefit of MIL is that it assigns labels to bags of instances instead of requiring a discrete label for each instance. Bags with a positive label contain at least one positive instance, while bags with a negative label contain only negative instances. The task of MIL is to learn the instances that contribute to a positively labeled bag; traditional MIL selects a single instance that is most representative of the entire bag via max-pooling. On the other hand, attention-pooling with MIL (AMIL) has shown to be effective at simultaneously considering the importance of each instance for classification[22].

We apply AMIL to the cytologic diagnosis of thyroid cancer from US images. Given biopsy reports associated with each patient, the final cytology of a set of US images for each patient can be used as a weak label for patient-level classification; therefore, all available data can be used for model development, including frames without nodule segmentations. We also explore the extraction of patches from US images of different sizes as input to the AMIL model, and fusing the pooled output of the model with clinical information, including age and sex. Patching removes the effect of pre-processing with different pad, rescale, or crop operations that could distort the underlying image information. Using patches as input to our AMIL models also offers the ability to interpret the models' decision by visualizing patch-level importance across multiple US images. Lastly, we analyze the utility of ensembling AMIL models trained on patches of different scales. Altogether, we demonstrate that the fusion of AMIL models trained on different scales of US image patches outperforms MIL models trained on features extracted from whole-image frames, as well as baseline machine-learning classification models trained on only clinical information.

Methods

Data

Table 1. Demographic information for samples with cytology labels (n=434).

Variable	Benign	Malignant
Total (N)	357	77
Sex (N,%)		
M	86 (24.0%)	24 (31.2%)
F	273 (76.0%)	53 (68.8%)
TI-RADS (N,%)		
1	9 (2.5%)	7 (9.1%)
2	25 (7.0%)	8 (10.4%)
3	135 (37.8%)	3 (3.9%)
4	170 (47.6%)	28 (36.36%)
5	18 (5.0%)	31 (40.3%)
Age (mean, std.)	57.35 ± 14.60	50.18 ± 17.14

All data was acquired with approval from the UCLA Institutional Review Board (IRB#19-001535). UCLA Health implements a standard protocol using different transducers at different locations to consistently achieve full coverage of the thyroid. Scans from patients undergoing this protocol have been aggregated into the UCLA Thyroid RadPath research dataset over the past thirteen years. The protocol yields 20-40 US images for each patient from different orientations and views. Cytology labels were extracted from the corresponding biopsy report for each patient. Any patients that had an indeterminate cytology label were excluded. For patients with multiple nodules at the time of US and biopsy, the cytology of the nodule with the highest malignancy risk was assigned as the label for the patient. Among all patients, labels were identified for 434 patients, of which 357 were benign and 77 were

malignant (Table 1) which provides 15,350 frames in total. Demographic information including sex, age, and TI-RADS score were also extracted. TI-RADS scores were missing for 49% of the patients due to acquisition prior to the implementation of the scoring system at UCLA Health. Missing scores were calculated based on TI-RADS associated key terms found in the corresponding radiology reports.

Baseline Model

Since the TI-RADS score is used to determine if further biopsy of a nodule is needed, it should hold some predictive power. Additionally, thyroid cancer tends to occur more often in women and at certain age ranges[23] and thus the sex and age of the patient should also hold predictive power. A logistic regression model was used to classify patients as benign or malignant using the values for sex, age, and TI-RADS score, which established a baseline classification performance using clinical and demographic data for later comparison against models trained using imaging data.

Multi-scale Imaging Model

The workflow for patient-level malignancy classification using US imaging data involves a series of steps (Figure 1). First, all the US images for a given patient were collected as a bag, and the bag was assigned a single patient-level label in accordance with the MIL framework. Next, lower-dimensional feature embeddings were extracted from either the entire US image frames, or patches of the US images. These features were passed through an AMIL model which outputted an aggregated feature vector weighted by the attentions allocated towards each instance. After calculating the aggregated feature representation for each patient, the feature vectors were passed through a classification layer to generate logits indicating the probability of malignancy. Prior to the classification layer, additional information such as clinical variables were optionally concatenated to the feature vector. The output logits from independently trained models could also be combined to enable the ensembling of different models trained on different scales of data, such as features from whole frames and patches of different dimensions. Given models that were independently trained on different scales of input data, the final classification probabilities could be averaged to ensemble the decision of each model; this method is more computationally feasible than training a single model on multiple scales of data, as well as forces each individual model to focus on relevant information available at each scale.

Whole Frame Feature Extraction

The extraction of lower-dimensional feature representations allows for smaller bags of feature instances to be passed to the AMIL model. To extract features from entire US image frames, we pre-processed the multiple images for a given patient, and passed each image through a deep feature extraction network; all the lower-dimensional embeddings for each patient were concatenated to form a consistent patient-level bag representation. Pre-processing of the US images ensured that the downstream feature extractor operated on consistently formatted images. Extraneous text details were removed from the left, right, top, and bottom borders of each image by cropping 125, 140, 100, and 100 pixels, respectively. The US probe orientation marker was removed using a 25px by 25px prototype marker to remove the patch that most likely contained the marker measured by pixel-wise difference. Intensity standardization was achieved by rescaling the intensity to the 2nd and 98th percentiles of intensity and then normalizing to zero mean and unit variance. All images were square-padded to 759px by 759px, which was the largest dimension across all images. To conform to the specifications of the feature extractor, the images were resized to 224px by 224px and duplicated along the 3 color channels. A convolutional neural network (CNN) was used to extract lower-dimensional feature representations of higher-dimensional images. We utilized an existing ResNet50 network architecture that was pre-trained on ImageNet to obtain meaningful features that represented the US images in a low-dimensional vector space[24]. Features were obtained from the output of the third hidden layer of the network, which consists of 1024 neurons. Concretely, each input image of dimensions $R^{224\times224\times3}$ was reduced to a $R^{1\times1024}$ vector. If n individual US images were associated with a single patient, then a bag representation of the patient was formed by concatenating the feature vectors to form a $R^{n\times1024}$ matrix for each patient.

Patching

Using a pre-trained feature extractor to obtain lower dimensional representations of each frame introduces several limitations, such as poor information representation after pre-processing and domain shift. Alternatively, we tiled each US image into uniform patches, and utilized a trainable CNN feature extractor to learn domain-specific features of non-distorted US data. The tiling process was parameterized by tile dimension and stride; we explored variations, including tiles of dimension 256px by 256px and 128px by 128px. Since each of the n individual US images associated with a single patient can be of a different size, each US image n_i can yield a variable number of patches p_i. The patches across all US images were concatenated to obtain p patches, forming a $R^{p\times m\times m}$ patch representation

for each patient, where m is the patch size. Passing the patch representation through a trainable feature extractor that reduced each m by m tile to a $R^{1\times1024}$ vector resulted in a final $R^{p\times1024}$ bag representation for each patient.

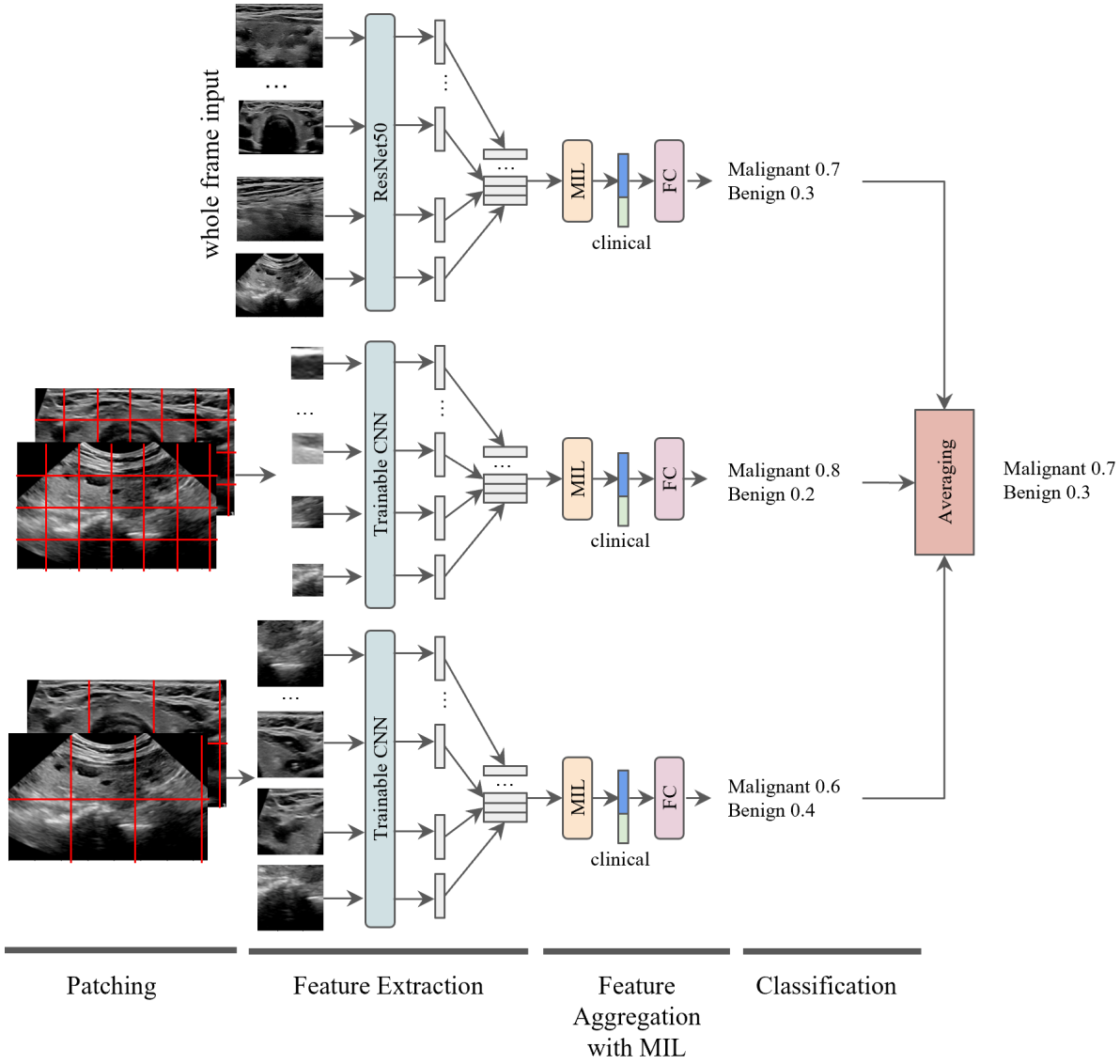

Figure 1. Attention-pooling multiple instance learning framework with multiple scales of data. The ensemble model consists of three parts: whole frame AMIL, patch AMIL with the patch size of 128px by 128px, and patch AMIL with a patch size of 256px by 256px. In the whole frame AMIL, each frame was passed into a pre-trained ResNet50 to obtain 1024-dimensional feature vectors. For the patch models, the whole frames were tiled into patches of two different sizes and then passed into trainable CNN for feature extraction. The extracted features were then concatenated and fed into an AMIL module to obtain a single aggregated feature vector for each patient. The clinical features can be appended to the aggregated vector, which is passed into a forward-connected layer for final malignancy prediction. Three models were trained independently. Two ensemble models were created by averaging the predicted probability from two patch models and all three models.

Multiple Instance Learning

Given bag representations of US imaging data for each patient, we utilized MIL to achieve thyroid nodule cytology classification[17,18]. MIL uses a single label to characterize an entire bag of instances, rather than requiring explicit information about each instance in a bag; concretely, individual US images or patches do not require independent labels. Instead, a bag is labeled as positive if it contains at least one single positive instance, and negative otherwise.

Therefore, the biopsy reports of a patient can be parsed to obtain a single label to characterize whether a set of patient's US images contains at least one malignant nodule (malignant) or not (benign). A report which mentions at least one malignant nodule in the patient's US images results in the set of images being labeled as malignant. Benign bags have no malignant nodules identified in the image set. This framework allows for the patient-level classification of thyroid malignancy in a set of US frames without requiring nodule annotations or labels for each frame or patch.

In the MIL framework, the model must learn the instances that contribute to the bag label. To achieve this goal, the model utilizes pooling functions after one or more learnable layers to aggregate the bag instances. We implemented AMIL to simultaneously consider the contribution of all instances towards the bag-level classification, rather than max-pooling which uses one instance to define a bag[22]. Each instance is assigned an attention score that defines the weighted contribution of the instance when aggregating into a single feature representation for bag-level classification. The attention scores are calculated with a gated attention layer constructed from two parallel streams of trainable linear layers with different activation functions[22].

Interpretability

In attention-based pooling, each instance is assigned an attention score representing weights when aggregating all instances into a bag-level classification; thus, attention-based pooling enables instance-level interpretability through directly informing which instances contribute the most towards the bag-level classification based on attention score. For bags consisting of whole-frame features, the attention scores indicate the relative importance of each frame towards the final contribution. For bags of patches, the attention scores reveal more granular information regarding which patches significantly contribute to the final classification. In both cases, the attention scores were normalized, mapped to a colormap, and reconstructed to visualize the highly-attended instances. For the AMIL models trained on patches, these visualizations were represented as heatmaps that highlight regions of interest (Figure 2). If the model classifies a patient as malignant, then the heatmaps indicate which regions the model weighed more when making that decision.

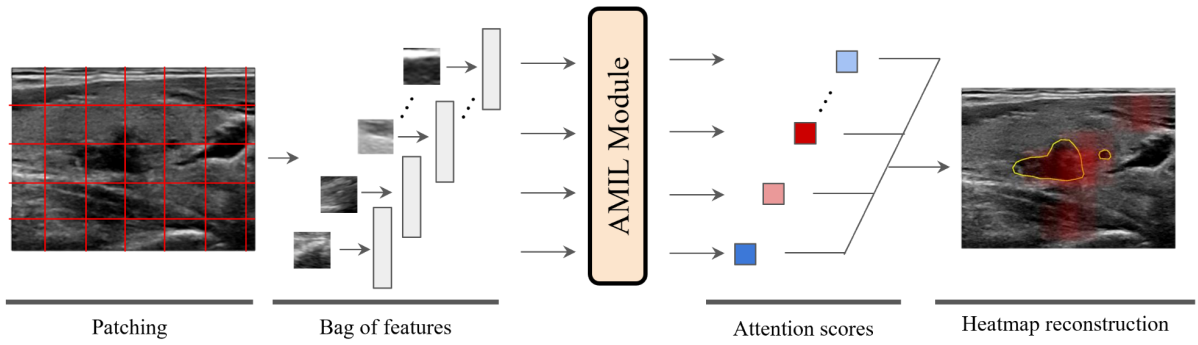

Figure 2. Heatmap interpretability pipeline from attention-based pooling with multiple-instance learning. AMIL assigns each tile with an attention score, which shows the importance of the tile toward malignancy prediction, and aggregates all patches based on the weights. The heatmap was reconstructed by extracting the attention scores from the AMIL module and mapping them back to the original whole frame. The high-attention regions were shown in red, and the nodule annotations obtained from radiologists were illustrated in yellow.

Experimentation and Evaluation

Patients were divided into five different non-overlapping splits of 80% training data and 20% validation data, stratified by the cytology labels, to achieve five-fold cross-validation. We trained different models using the same five-fold splits. We compared three different patient bag representations: whole-frame features from ResNet50, patch features from 256px by 256px patches, and patch features from 128px by 128px patches. We also combined the two best models trained on the different scales of patch features and averaged the output probabilities. Lastly, we ensembled the best models trained on the patch features and whole-frame features to obtain the final combined configuration. In total, the best configurations of these five primary model variations are reported in the results.

For each of the above model variations, we retrained on a grid of hyperparameters to determine the optimal configurations. However, it is essential to underscore that we maintained the default values for all training-related hyperparameters (such as learning rate and batch size) as well as model-specific hyperparameters (such as model architecture) throughout the entire experimentation process. The trainable CNN that extracts features from patches was constructed with and without Dropout layers between each convolutional layer. Prior to the final classification layer, we either passed the aggregated features or concatenated additional clinical variables (age, sex). To train the classification task, we compared both weighted cross-entropy loss and focal loss, to account for the class imbalance[25]. Lastly, we attempted to augment the training data with random horizontal flips and random rotations between -15 and 15 degrees.

Each model was trained with early-stopping based on minimum validation loss with a patience of 10 epochs. The probability of predicting each patient bag in the validation set as malignant was compared against each respective true label. The area under the receiver operating characteristic curve (AUROC) was computed to evaluate the classification performance of the models. Given that the cytology labels are imbalanced, the area under the precision-recall curve (AUPRC) was also used to better assess the performance of the classifier on the minority class[26]. We also reported the F1 score at the threshold that yielded the maximum F1 score, and the corresponding accuracy, precision, and recall at the same threshold. The statistical significance of proposed models relative to the baseline was evaluated at a significance level of 0.05 using the DeLong Test for AUROC and Wilcoxon signed-rank test for accuracy[27]. We also used the Benjamini-Hochberg procedure to adjust the p-values to correct for false-discovery rate.

Results

We developed and fine-tuned three models: frame AMIL, patch (256px by 256px) AMIL, and patch (128px by 128px) AMIL. In addition, we evaluated two ensemble models that combined the two patch models and all three models, respectively. To assess the model performance, the mean and standard deviation of AUROC, AUPRC, F1, accuracy, precision, and recall across five folds from cross-validation were computed. The detailed experimental results of models that yield the best AUPRC are presented in Table 2. The predicted probabilities of malignancy across five folds were combined together to plot the ROC and PR curves of all models, shown in Figure 3.

Table 2. Model performance. Mean and standard deviation of AUROC, AUPRC, F1, accuracy, precision, and recall across five folds (*statistically significant at a significance level of 0.05).

Models	AUROC	AUPRC	F1	Accuracy	Precision	Recall
baseline	0.667±0.103	0.444±0.133	0.499±0.139	0.738±0.190	0.460±0.182	**0.658±0.221**
frame	0.636±0.038	0.363±0.054	0.406±0.047	0.689±0.124	0.351±0.122	0.572±0.141
patch (128)	0.742±0.101	0.502±0.141	0.515±0.116	0.788±0.104*	**0.516±0.242**	0.583±0.108
patch (256)	0.737±0.075	0.443±0.083	0.495±0.096	0.765±0.077	0.424±0.123	0.637±0.152
ensemble (patch 128 and 256)	**0.785±0.104***	0.539±0.146	**0.541±0.108**	**0.813±0.062***	0.502±0.122	0.621±0.185
ensemble (patch 128 and 256 and frame)	0.772±0.098*	**0.548±0.121**	0.531±0.091	0.800±0.066*	0.480±0.134	0.623±0.118

The baseline model trained on sex, age, and TI-RADS score achieved an accuracy of 0.738, AUROC of 0.667, AUPRC of 0.444, and F1 score of 0.499. The coefficients of the logistic regression model indicated that the TI-RADS score (0.38) and age (-0.48) are more important features than sex (-0.20). Although the model trained on the whole frame features did not perform as well as the baseline model, the models independently trained on two patch sizes outperformed the baseline model. Specifically, the model trained on 256px by 256px patches achieved a higher AUROC (0.737), and the model trained on 128px by 128px patches surpassed the baseline in terms of AUROC (0.742), AUPRC (0.502), F1 score (0.515), and accuracy (0.788, $p = 0.045$). Additionally, the ensemble model that aggregates the predictions from patch models trained on 128px and 256px achieved the highest performance in all evaluation metrics with an accuracy of 0.813 ($p = 0.008$), AUROC of 0.785 ($p = 0.024$), AUPRC of 0.539, and F1 score of 0.541. Although the ensemble of the frame and two patch models still presented promising results (AUROC: 0.772, $p = 0.025$; AUPRC: 0.548; F1: 0.531; Accuracy: 0.800, $p = 0.017$) compared to the baseline, the inclusion of frame model did not lead to further improvement in the performance of the ensemble

model. Furthermore, Figure 3 highlights the ROC and PR curves that demonstrate that the two ensemble models achieved the best performance.

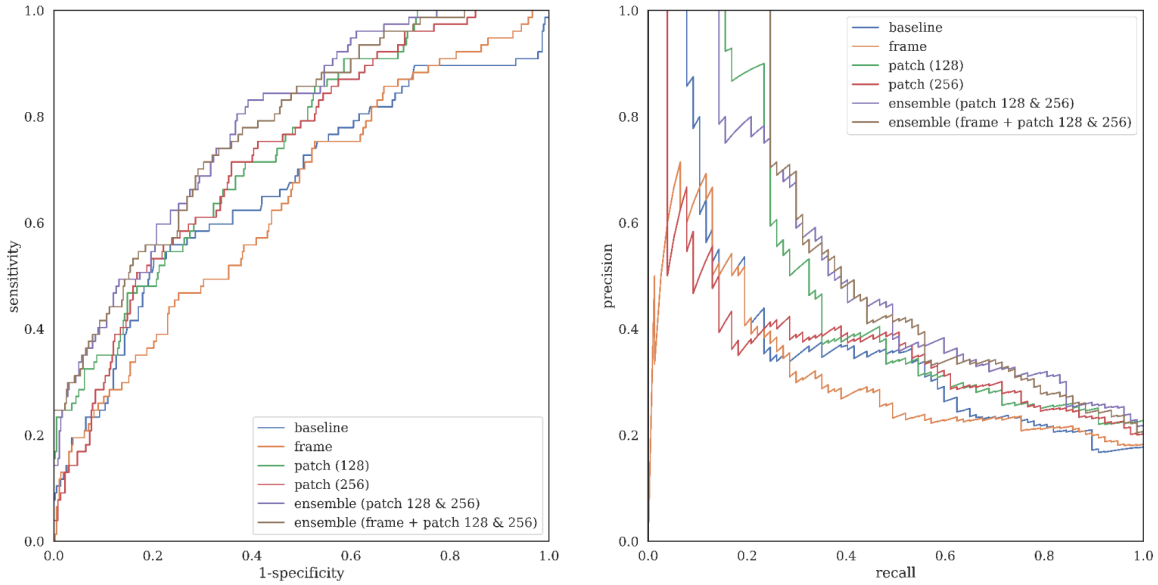

Figure 3. ROC and PR curves for all models. Predicted probabilities across five folds of each model were combined together to create the ROC (left) and PR (right) curves. The two ensemble models demonstrate the highest area under the curve in both ROC and PR curves.

We utilized attention scores of two distinct AMIL models, each trained on patches with 128px by 128px and 256 px by 256px, to generate a heatmap for each frame. Examples of the resulting heatmaps for a single patient with malignant nodules are displayed in Figure 4. The generated heatmaps highlight the region with high-level attention in red which corresponds to the area which contains the nodules.

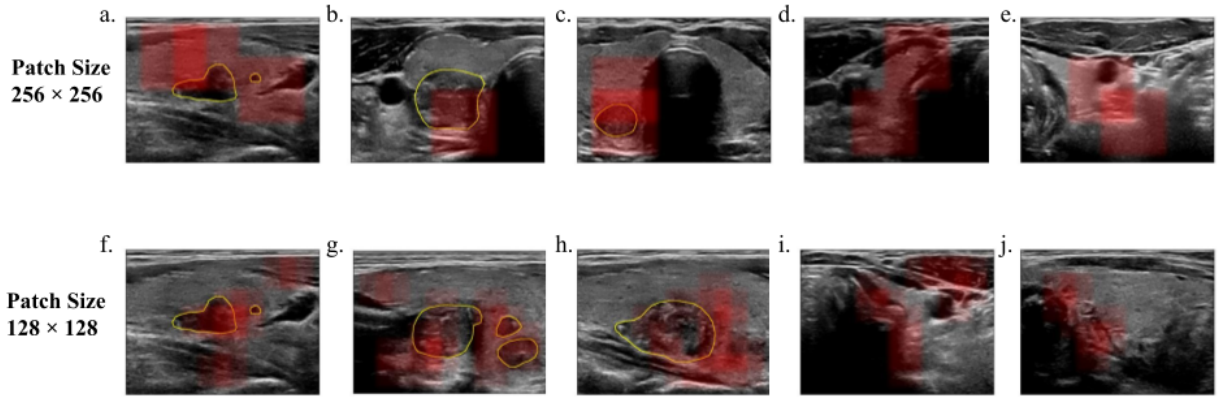

Figure 4. Attention heatmap for multiple US frames of a single, correctly-classified, malignant patient. The level of attention can be inferred by the degree of redness exhibited in the image, and the nodule annotations obtained from radiologists were illustrated in yellow. The models pay high attention to the nodule, especially along the edges of the nodules (a-c; f-h), and also to the non-nodule hyperechoic areas (d-e; i-j).

Discussion and Conclusions

Overall, the results demonstrate that incorporating US images within an AMIL workflow improves the classification performance of patient-level malignancy compared to baseline performance when using information from radiology reports. When using only features from whole US frames, performance is worse than baseline, which suggests that the pipeline involving whole-frame image pre-processing and a ResNet50 feature extractor does not yield

comprehensive representations of the original images. On the other hand, the use of patch features results in improvements in classification performance, indicating that patch features are more representative of the original image information. Models trained using features from 128px by 128px patches have a more balanced precision and recall at the maximum F1 score, while models trained using features from 256px by 256px patches are biased towards recall. The results from the ensembled models that average the logit probabilities from individually trained models result in significant improvement measured by AUROC. Both cases where the outputs of models trained on different patch scales are averaged, as well as different patch scales and whole US frames, result in significant improvement. Thus, we achieve the best performance when combining models that are independently trained on different scales of information. Whole-frame features do not contribute to improvement in performance, suggesting that a multi-scale approach using patch data contains the best information when assessing the malignancy of thyroid nodules.

The heatmap visualizations show that the models pay high attention to nodule regions that are clinically relevant for diagnosis; however, several of the images show a mix of attention towards nodule regions, nodule edges, and non-nodule regions. Consultations with radiologists reveal that the attended areas, including false-positive regions, correspond to calcifications along the edges of nodules, irregular edges, as well as nodule and non-nodule hyperechoic areas. Radiologists use similar information when scoring TI-RADS, indicating that our models are paying attention to clinically relevant regions.

Previous work in the literature achieves exceptional thyroid nodule malignancy classification by applying deep convolutional neural networks for feature extraction and classification[11,12,13,14,15]; however, these studies focus on the classification of individual nodules, whereas we pursue a more realistic approach of patient-level classification. Furthermore, they all depend on manual annotations of the nodules, while we develop methods that are not as dependent on such labeled data. Other work approaches the problem using a similar MIL approach as ours; however, such work utilizes additional elastograms, as well as hand-annotated lesions[28]. Another approach is similar to our end goal, but the method limits the number of images utilized for each patient, as well as evaluates on radiology results rather than cytology results[29]. None of the existing work utilizes different scales of information. We contribute a model that achieves patient-level thyroid malignancy classification without requiring annotated data, using AMIL with multiple scales of input US image features.

One limitation of our work is the use of a dataset from a single institution without external validation. Future work includes the curation of external data to validate our workflow. Secondly, the US images are captured at different views resulting in intrinsic variations in scale of the field of view. Such information is relevant because consistent patch sizes in image space may not correspond to equal sizes in true space. Future work can involve obtaining such information and assessing the impact on models that have been calibrated to such variations.

In conclusion, we have developed a multi-scale AMIL pipeline that performs patient-level malignancy classification from multiple thyroid US images per patient. In addition to mitigating the dependency on pre-processing and pre-trained feature extractors, patching helped focus on different scales of data, as well as refine the interpretability from frame-level attention to patch-level attention. We demonstrate improvement over a baseline model that uses TI-RADS data. The clinical application of an improved model can utilize the automatic analysis of imaging data to better discern risk of malignancy, especially for benign cases, and ultimately achieve the goal of reducing unnecessary biopsies.

References

1. Wong R, Farrell SG, Grossmann M. Thyroid nodules: diagnosis and management. Medical Journal of Australia.2018;209(2):92–8.
2. Abdolali F, Kapur J, Jaremko JL, Noga M, Hareendranathan AR, Punithakumar K. Automated thyroid nodule detection from ultrasound imaging using deep convolutional neural networks. Computers in Biology and Medicine. 2020 Jul 1;122:103871.
3. Rago T, Vitti P. Role of thyroid ultrasound in the diagnostic evaluation of thyroid nodules. Best Practice & Research Clinical Endocrinology & Metabolism. 2008 Dec 1;22(6):913–28.
4. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. Journal of the American College of Radiology. 2017 May 1;14(5):587–95.

5. Amedee RG, Dhurandhar NR. Fine-Needle Aspiration Biopsy. The Laryngoscope. 2001;111(9):1551–7.
6. Alshaikh S, Harb Z, Aljufairi E, Almahari SA. Classification of thyroid fine-needle aspiration cytology into Bethesda categories: An institutional experience and review of the literature. Cytojournal. 2018 Feb 16;15:4.
7. Papini E, Guglielmi R, Bianchini A, Crescenzi A, Taccogna S, Nardi F, et al. Risk of Malignancy in Nonpalpable Thyroid Nodules: Predictive Value of Ultrasound and Color-Doppler Features. The Journal of Clinical Endocrinology & Metabolism. 2002 May 1;87(5):1941–6.
8. Mandel SJ. Diagnostic Use of Ultrasonography in Patients with Nodular Thyroid Disease. Endocrine Practice. 2004 May 1;10(3):246–52.
9. Chi J, Walia E, Babyn P, Wang J, Groot G, Eramian M. Thyroid Nodule Classification in Ultrasound Images by Fine-Tuning Deep Convolutional Neural Network. J Digit Imaging. 2017 Aug 1;30(4):477–86.
10. Nguyen DT, Choi J, Park KR. Thyroid Nodule Segmentation in Ultrasound Image Based on Information Fusion of Suggestion and Enhancement Networks. Mathematics. 2022 Jan;10(19):3484.
11. Kwon SW, Choi IJ, Kang JY, Jang WI, Lee GH, Lee MC. Ultrasonographic Thyroid Nodule Classification Using a Deep Convolutional Neural Network with Surgical Pathology. J Digit Imaging. 2020 Oct;33(5):1202–8.
12. Yang J, Shi X, Wang B, Qiu W, Tian G, Wang X, Wang P, Yang J. Ultrasound image classification of thyroid nodules based on deep learning. Frontiers in Oncology. 2022 Jul 15:3545.
13. Zhang C, Liu D, Huang L, Zhao Y, Chen L, Guo Y. Classification of Thyroid Nodules by Using Deep Learning Radiomics Based on Ultrasound Dynamic Video. Journal of Ultrasound in Medicine. 2022;41(12):2993–3002.
14. Wang J, Li S, Song W, Qin H, Zhang B, Hao A. Learning from Weakly-Labeled Clinical Data for Automatic Thyroid Nodule Classification in Ultrasound Images. In: 2018 25th IEEE International Conference on Image Processing (ICIP). 2018. p. 3114–8.
15. Chi J, Walia E, Babyn P, Wang J, Groot G, Eramian M. Thyroid Nodule Classification in Ultrasound Images by Fine-Tuning Deep Convolutional Neural Network. J Digit Imaging. 2017 Aug 1;30(4):477–86.
16. Wang L, Zhou X, Nie X, Lin X, Li J, Zheng H, Xue E, Chen S, Chen C, Du M, Tong T. A Multi-Scale Densely Connected Convolutional Neural Network for Automated Thyroid Nodule Classification. Frontiers in Neuroscience. 2022;16.
17. Yang J. Review of multi-instance learning and its applications. Technical report, School of Computer Science Carnegie Mellon University. 2005.
18. Maron O, Lozano-Perez T. A Framework for Multiple-Instance Learning.
19. Sudharshan PJ, Petitjean C, Spanhol F, Oliveira LE, Heutte L, Honeine P. Multiple instance learning for histopathological breast cancer image classification. Expert Systems with Applications. 2019 Mar 1;117:103–11.
20. Kandemir M, Feuchtinger A, Walch A, Hamprecht FA. Digital pathology: Multiple instance learning can detect Barrett's cancer. In: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI). 2014. p. 1348–51.
21. Chen J, Zeng H, Zhang C, Shi Z, Dekker A, Wee L, et al. Lung cancer diagnosis using deep attention-based multiple instance learning and radiomics. Medical Physics. 2022;49(5):3134–43.
22. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. InInternational conference on machine learning 2018 Jul 3 (pp. 2127-2136). PMLR.
23. Thyroid cancer – trends by sex, age and histological type [Internet]. [cited 2023 Mar 2]. Available from: http://www.ncin.org.uk/publications/data_briefings/thyroid_cancer_trends_by_sex_age_and_histological_type
24. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).
25. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. InProceedings of the IEEE international conference on computer vision 2017 (pp. 2980-2988).
26. Ma Y, He H, editors. Imbalanced learning: foundations, algorithms, and applications.
27. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988;44(3):837–45.
28. Ding J, Cheng HD, Huang J, Zhang Y. Multiple-instance learning with global and local features for thyroid ultrasound image classification. In2014 7th International Conference on Biomedical Engineering and Informatics 2014 Oct 14 (pp. 66-70). IEEE.
29. Wang L, Zhang L, Zhu M, Qi X, Yi Z. Automatic diagnosis for thyroid nodules in ultrasound images by deep neural networks. Medical Image Analysis. 2020 Apr 1;61:101665.

Supplemental Information

Table 4. Hyperparameters used for the frame and patch AMIL models. We trained our models with hyperparameters: loss functions (focal loss or weighted cross-entropy loss), augmentation (no or yes), drop-out layers in trainable CNN (added or not added), and clinical features (included or not included). Here, we show the hyperparameters we used for the model with the best AUPRC score.

Models	Loss Function	Augmentation	Dropout	Clinical Features	Stride size
frame	Focal Loss	Yes	-	Yes	-
patch (128)	Weighted Cross-Entropy Loss	Yes	No	No	64
patch (256)	Focal Loss	No	Yes	Yes	256

Table 5. Performance of a late fusion model. Rather than ensembling models through the averaging of output probabilities, we also attempt an alternative approach of late fusion where aggregated feature representations are concatenated prior to the final classification layer. We document the mean and standard deviation of AUROC, AUPRC, F1, precision, and recall across five folds. Performance does not significantly improve upon baseline performance, in addition to significantly slower training speed (Table 5).

Models	AUROC	AUPRC	F1	Precision	Recall
Late fusion	0.685 ± 0.102	0.450 ± 0.095	0.457 ± 0.105	0.489 ± 0.221	0.546 ± 0.223