Automatic Detection of Thyroid Nodule Characteristics from 2D Ultrasound Images

Dongxu Han[1], Nasir Ibrahim[1], Feng Lu[2], Yicheng Zhu[3], Hongbo Du[1] and Alaa AlZoubi[4]

Dongxu Han[1] dongxu.han@buckingham.ac.uk

Nasir Ibrahim[1] nasir.ibrahim@buckingham.ac.uk

Feng Lu[2]: uslufeng@163.com

Yicheng Zhu[3] ycheng_zmd@126.com

Hongbo Du[1] hongbo.du@buckingham.ac.uk

Alaa AlZoubi[4] a.alzoubi@derby.ac.uk

[1] School of Computing, the University of Buckingham, Buckingham, MK18 1EG, United Kingdom

[2] Department of Medical Ultrasound, Shanghai Tenth People's Hospital, School of Medicine, Tongji University, Shanghai, 200072, China

[3] Department of Ultrasound, Pudong New Area People's Hospital affiliated to Shanghai University of Medicine and Health Sciences, Shanghai, 201200, China

[4] School of Computing and Engineering, University of Derby, Derby, DE22 3AW, United Kingdom

Abstract

Purpose: Thyroid cancer is one of the common types of cancer worldwide, and Ultrasound (US) imaging is a modality normally used for thyroid cancer diagnostics. The American College of Radiology Thyroid Imaging Reporting and Data System (ACR TIRADS) has been widely adopted to identify and classify US image characteristics for thyroid nodules. This paper presents novel methods for detecting the characteristic descriptors derived from TIRADS.

Methods: Our methods return descriptions of the nodule margin irregularity, margin smoothness, calcification as well as shape and echogenicity using conventional computer vision and deep learning techniques. We evaluate our methods using datasets of 471 US images of thyroid nodules acquired from US machines of different makes and labelled by multiple radiologists.

Results: The proposed methods achieved overall accuracies of 88.00%, 93.18% and 89.13% in classifying nodule calcification, margin irregularity, and margin smoothness respectively. Further tests with limited data also show a promising overall accuracy of 90.60% for echogenicity and 100.00% for nodule shape.

Conclusion: This study provides an automated annotation of thyroid nodule characteristics from 2D ultrasound images. The experimental results showed promising performance of our methods for thyroid nodule analysis**.** The automatic detection of correct characteristics not only offers supporting evidence for diagnosis, but also generates patient reports rapidly, thereby decreasing the workload of radiologists and enhancing productivity.

Keywords

Thyroid cancer, Ultrasonography, TIRADS, Nodule characteristics, Machine learning, Computer-aided Diagnosis

1. Introduction

Thyroid cancer is one of the most lethal cancers globally [1]. The incidence rate in women is three times higher than that in men; in 2018 alone, one in 20 women diagnosed of cancer had thyroid cancer [1]. Different imagery systems have been used for diagnosis. US imaging has the advantages of being non-invasive, non-radiative and of low-cost. However, recognising thyroid nodule and detecting cancer characteristics from US images are challenging due to the demanding skills required in image acquisition and low image quality caused by speckle noise and artefacts. To tackle the issues and maintain consistency in clinical settings, doctors often use standard guidelines in describing thyroid nodules. The original TIRADS principles were first proposed in [15]. It was later standardized by Kwak, et al [2] as the first reporting scheme for classifying thyroid nodules to risk levels of malignancy using US nodule characteristic descriptors. The most recent ACR TIRADS further standardizes the descriptors to five categories [3]. Although radiologists have been using the guidelines to report thyroid nodules under different conditions, accurate diagnosis based on TIRADS remains challenging because of inter- and intra-observer variabilities.

Several studies have been conducted to analyse US image characteristics of thyroid nodules, but most of them extracted such characteristics for classifying a nodule as benign or malignant rather than accurately detecting and evaluating the characteristics for report generation [6, 9, 10]. In this paper, we present a comprehensive translation of the US TIRADS characteristics, aiming at an automated process of describing clinical findings in thyroid nodules. The proposed methods provide an effective, efficient, deterministic, and consistent annotation of the TIRADS terms to reduce subjectivity and increase precision in nodule examination and reporting. In particular, the paper is intended to make the following key contributions: (1) a new method for nodule irregularity detection by utilizing convexity, ellipticity, lobulation, and angulation features; (2) a new method for smoothness detection using texture features and super-pixels; (3) an optimized CNN classification model (CaNet) for calcification by using Bayesian Optimization; and (4) a new method that combines CaNet and super-pixels for more accurate calcification classification.

The remaining part of the paper is organised as follows. Section 2 reports on the key TIRADS US characteristics and reviews existing methods for nodule characteristics analysis in US images. Section 3 presents the proposed methods for detecting margin irregularity, margin smoothness and calcification. Section 4 evaluates the methods through experiments on datasets collected from clinics. Section 5 further discusses possible methods for detecting nodule shapes and echogenicity before Section 6 summarizes the main findings and concludes the paper.

2. Background and Related Work

The ACR TIRADS scheme characterizes a thyroid nodule from five aspects: composition, echogenicity, shape, margin, and echogenic foci [3]. Each aspect contains a set of descriptive terms with associated points. Based on the observation, the associated points are added to a total score and then mapped to one of five ordinal bands (from TR1 to TR5). A benign nodule within the TR2 category can exhibit a regular oval shape, an anteroposterior transverse ratio (AP/T ratio) that is wider than tall, a smooth margin, anechoic properties, and the absence of calcification. On the other hand, a malignant nodule falling under the TR5 category may show a lobulated shape with an irregular margin, an AP/T ratio that is taller than wide, hypo-echogenicity, and the presence of micro-calcification. The borderline TR4 band is further divided into 4a, 4b and 4c sub-bands. It is also the band where

most inter-observer variability occurs, and hence there is a greater need for nodule characteristics for correct decisions. Besides the ACR TIRADS, other guidelines also exist [11, 12,13], all of which indicate similar nodule characteristics of suspected malignancy [14]. Therefore, in practice, hospitals normally use common categories of thyroid nodule characteristics across the different guidelines. It must be said that a TIRADS score is still an observer's subjective judgement and may lead to different diagnosis outcomes. Rigorous definitions of the terms and reviews of the guidelines may help reducing but not avoiding such variability. Having an automated computer-based solution to categorize the thyroid nodules may help reducing such non-deterministic outcomes. Using the detected characteristics as evidence will enhance comprehensibility of the final diagnostic decisions and gain trust from the medical profession.

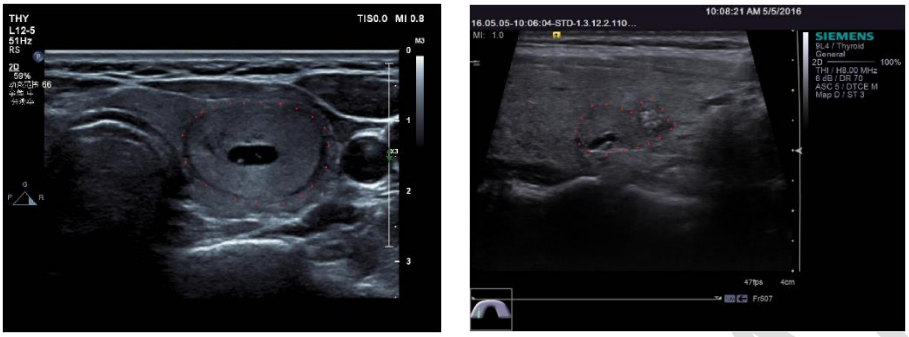
Several studies to automatically quantify features based on the standardized TIRADS categories for classifying thyroid nodules have been reported [4, 5], but the work on extracting correct TIRADS features from ultrasound images for annotation purposes remains limited. Zulfanahri et al [6] analysed and classified the margin irregularity of thyroid nodule into regular or irregular class using rectangularity, convexity, and tortuosity features with an SVM classifier. The study reported an accuracy of 91.52% (91.80% sensitivity and 91.35% specificity) over a set of 165 images. Wang et al [4] automated the extraction of four thyroid nodule characteristics: composition using average image brightness, echogenicity pattern using relative brightness, calcification with top-hat morphological filter, and boundary regularity with acutance. The study used a semi-supervised fuzzy C-means ensemble (SS-FCME) model to classify the thyroid nodules into a TIRADS score band with 70.77% overall accuracy. Nugroho et al [10] trained an SVM to determine the margin of a thyroid nodule using compactness, convexity, circularity, dispersion, aspect ratio, rectangularity, solidity and tortuosity as features, with an accuracy of 92.30% on 144 test images. In a later study [9], Nugroho et. al. further added the orientation feature and achieved an accuracy of 98.00% but only on 51 test images. Although the previous two studies reported promising performances, the features extracted can be too excessive for the problems. Zhuang et al [7] used cystic growth rate and the variance of the grayscale distribution for composition, compactness for margin irregularity, and the aspect ratio for nodule shape. A deep learning algorithm was used to classify calcification based on ROI image, but the paper lacks detailed explanations. For margin smoothness, the method first locates a 10-pixel ribbon around the nodule boundary (inside and outside regions) via morphological dilation and erosion. Average grayscale difference (or mean separability), derived from the number of pixels, mean and variance of intensity in each region, was used to quantify margin smoothness. Weights were assigned to the derived quantities and feature scores that were then accumulated to the total TIRADS score. All malignant cases and 94.87% of the benign cases were classified into the correct TIRADS score bands.

3. Materials and Methods

3.1. Dataset Collection and Annotation for Nodule Characteristics Analysis

Thyroid cancer has various subtypes. Malignant tumours have more diversity in their cellular and molecular structures than benign tumours. Therefore, including a larger number of malignant cases in a dataset is important to ensure sufficient representation of the diverse subtypes for developing accurate models. Hence we purposely included more malignant cases of various pathologies in this study. During the image acquisition, one radiologist with more than 15 years of experience manually cropped every nodule in each original image by identifying coordinate points on the nodule boundary. The delaminated boundaries were further validated by the second

radiologist with a similar amount of experience. Images with disagreed nodule boundaries were excluded from the final data collection. The verified nodule boundaries form polygons for the Region Of Interest (ROI). Fig.1 shows two examples from the dataset.



(a) (b)

Fig. 1 Example US Images with labelled ROI (red dots on nodule boundaries): (a) isoechoic/hyperechoic, wider-than-tall, clear, regular and no calcification, and (b) isoechoic/hyperechoic, wider-than-tall, not clear, irregular, micro calcification

Our dataset was labelled by 3 radiologists with 10, 15 and 30 years of experience respectively. For each ROI identified, US image descriptors of the nodule were assigned by at least 2 radiologists following the ACR TI-RADS guideline, one of whom is ensured to have at least 15 years of experience. When labelling the margin and echogenicity foci, the radiologists made modifications to align with the current clinical practices and reduce the observer viabilities. In particular, the classification of margin is refined into two subcategories as margin irregularity (irregular or regular) and margin smoothness (not-clear or clear). The echogenic foci is simplified as "no calcification", "microcalcification" or "macrocalcification". Nodules containing both micro and macro calcifications is classified as microcalcification as it indicates a higher risk of malignancy. In the end, a dataset of total 471 thyroid ultrasound images from two local hospitals in Shanghai, China was obtained and labelled. The dataset contains 140 benign cases and 331 malignant cases, where the pathology result of each image was confirmed by FNA. All the personal details of the patients were excluded. The collected dataset was randomly divided into three equal size patches (157 each) respectively for training, validation and testing purposes (to be further explained in the next section). Both agreed and disagreed labels are recorded without additional bias for performing multi-observer studies in the later experiment.

3.2. Nodule Characteristics Detection Methods

Despite some literature suggestions at possible transferring learning when analysing thyroid and breast tumours as they are both superficial organs [17], it may be rather difficult to adapt models from other types of organs directly for characteristics analysis for thyroid nodules as they share different definitions. Some characteristics such as calcification can be difficult to analyse for breast tumours due to the limitation of ultrasonography [20]. Nevertheless, some characteristics such as margins do share similarities when used for classifying cancer malignancy [21], but their characteristics can still vary as they are growing on different mediums with possibly different cell types. Therefore, we have proposed a list of novel methods for analysing thyroid characteristics with insights from existing literatures.

Margin Irregularity

Margin irregularity is a characteristic that studies the geometric shape of the nodule margin. Fig.2 shows an example of a nodule with an irregular margin. Based on the nodule boundary coordinates, the algorithm derives the convex variance, elliptic variance, lobulation, and angulation from the ROI margin; each of them captures a unique feature for measuring margin irregularity and contributes to a final decision.

Irregularity Measure Extraction: Margin irregularity is first analysed by lobularity. Given a set of concave regions $\{c_1, c_2, \dots, c_n\} \in \mathcal{C}$

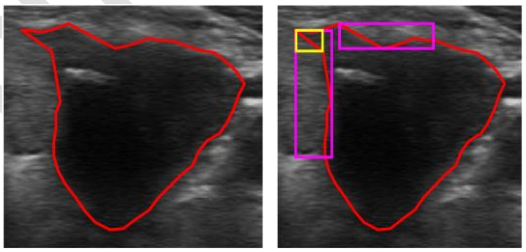
$$f_L(c) \coloneqq \begin{cases} \text{Lobular}, & if\ A_C \geq t_A \wedge \min(w_C, h_C) \geq t_l \\ \text{Not Lobular}, & else \end{cases} \tag{1}$$

where A_C denotes the ratio of the area of the concave region to that of the entire nodule, w_C and h_C denote the width and height of the concave region, t_A and t_l are the two thresholds defined for classification and determined empirically as 0.015 and 5 respectively.

Angulation is also an important factor for margin irregularity. Angulation analysis focuses on the extension around the margin. Therefore, the algorithm examines the spikiness, roughness and distortions of the margin by using a set of three consecutive coordinates, $\{p_1, p_2, p_3\}, \{p_2, p_3, p_4\}, \dots, \{p_{|\mathcal{P}|-1}, p_{|\mathcal{P}|}, p_{|\mathcal{P}|+1}\}$, from the total set of ROI margin coordinates $\mathcal{P}$ where $p_0 = p_{|\mathcal{P}|}$ and $p_{|\mathcal{P}|+1} = p_1$. The algorithm measures the curvature κ of the coordinates $\{p_{n-1}, p_n, p_{n+1}\}$. A large amount indicates a sharp change on the margin at the given coordinates [10]. The algorithm also calculates the angle θ: $p_{n-1} \rightarrow p_n \rightarrow p_{n+1}$, where a large angle indicates a slow change and a small angle indicates a sharp change at the given coordinates. The κ and θ values are then combined to estimate the angulation using the rule f_A in Eq. 2:

$$f_A(p_{n-1}, p_n, p_{n+1}) \coloneqq \begin{cases} \text{Angular}, & if\ \kappa \geq t_\kappa \wedge \theta < t_\theta \\ \text{Not Angular}, & else \end{cases} \tag{2}$$

where t_κ and t_θ are two thresholds used for classification and determined empirically as 0.1 and 90° respectively.



(a) (b)

Fig. 2 Illustration of Margin Irregularity Detection (a) Nodule ROI with coordinates (red) (b) An irregular nodule with lobulation (magenta) and angular (yellow) regions.

Irregularity Classification: With the four irregularity measures determined, the margin irregularity of a nodule is classified using the rule in Eq. 3:

$$\begin{cases} \text{Regular}, & if\ \ \sigma_c^2 \geq t_C \vee \sigma_e^2 \geq t_E \wedge f_L(\mathcal{C}) + f_A(\mathcal{P}) = 0 \\ \text{Irregular}, & else \end{cases} \tag{3}$$

where

$$f_L(\mathcal{C}) = \sum_{j=1}^{|\mathcal{C}|} \mathbf{1}_{\{\text{Lobular}\}}[f_L(c_j)],\ f_A(\mathcal{P}) = \sum_{k=1}^{|\mathcal{P}|} \mathbf{1}_{\{\text{Angular}\}}[f_A(p_{k-1}, p_k, p_{k+1})]$$

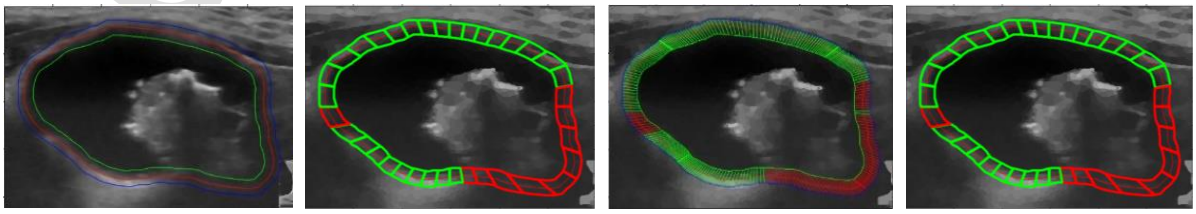
and t_C and t_E are two thresholds for classification, which are determined experimentally as 0.9 and 0.97 respectively.

Our irregularity method and the methods developed in [4, 6, 7, 10] both measure global irregularity of the nodule, but our method uses convexity and ellipticity variance, providing a more robust and accurate assessment of nodule irregularity without being excessive. Furthermore, our method has a new feature extraction step that incorporates and measures of local irregularity of margin such as lobulation and angulation. This feature extraction step has shown an improved sensitivity of margin irregularity demonstrating the better effectiveness of our method.

Margin Smoothness

Margin Smoothness represents the clarity of the nodule margin that is reflected the intensity contrast. The higher the contrast between regions inside the nodule boundary and the regions outside, the clearer the margin is. For better results, we first pre-process the US images to suppress the speckle noise and enhance the images. An adaptive median filter [11] is applied first for reducing the noise. This is then followed by bilateral filtering with Gaussian kernels [12] to enhance edges. The pre-processed image is then masked by the ROI delimitated by the radiologists to determine inside and outside ribbons around the nodule margin with the assistance of morphological operations (see Fig.3a). These ribbons are further divided into $|R|$ equal regions of $\frac{360°}{|R|}$ each, where R can be determined heuristically depending on the trade-off between precision and computation cost. In this study, $|R|$ is set to 36 (see Fig.3b), allowing more precise estimations of local margin smoothness.

Smoothness Measure Extraction: To measure margin smoothness, we first represent each region of the inner and outer ribbon with a two-dimensional vector composed by the averages and variances of the pixel intensities within the region. The difference between each pair of inner and outer regions is measured using Euclidean distance, where higher difference implies a clearer margin. It is noted that the proposed measure can only represent a general smoothness over a whole region due to its statistical nature. It can be a drawback especially when analysing small lesions. To overcome this drawback, we derive another measure that examines intensity profiles across the inner and outer ribbons at a 2° interval within each region, focusing on changes in fine details. Since such an intensity profile can be sensitive to noises, we have further processed the ultrasound image into superpixels using the SLIC algorithm [13]. Each profile is considered *distinct* if the difference between the highest and lowest superpixel readings is greater than t_λ (defined empirically as 20); otherwise, *indistinct* (see Fig.3c). The region outputs and intensity profile outputs per region are used to derive the final distinctiveness for that region (see Fig.3d).



(a) (b) (c) (d)

Fig. 3 Illustration of Margin Smoothness Detection. (a) Nodule ROI with outer (blue) / original (red) / inner (green) ribbons, (b) global detection results (distinct (green) / indistinct (red)), (c) local detection results (distinct (green) / indistinct (red)), (d) final smoothness prediction (distinct (green) / indistinct (red))

Smoothness Classification: The algorithm fuses the decision for the region analysis (RD), Fig 3b, and the decision for the signal analysis of intensity profiles (SD), Fig. 3c, in classifying the margin smoothness for each region. Each region's smoothness, SM_r, is classified as clear or not clear using Equ. 4.

$$SM_r = \begin{cases} \text{Clear}, & if \quad RD_r = Disinct \wedge SD_r \geq t_p \\ \text{Not Clear}, & else \end{cases} \tag{4}$$

where, r is a region between 1 .. |R|, $SD_r = \frac{\sum_{i=1}^{|S|} IPD_i}{|S|}$, IPD $(Intensity\ Profile\ Decision)$ is the decision for each intensity profile within the region r, $|S|$ is the number of intensity profiles in region r and t_p is the intensity profile classification threshold experimentally as 80%. The overall margin smoothness of a nodule is classified using Equ. 5.

$$Smoothness, SM = \begin{cases} \text{Clear}, & if \quad \frac{\sum_{r=1}^{|R|} SM_r}{|R|} \geq t_S \\ \text{Not Clear}, & else \end{cases} \tag{5}$$

where, t_s is the classification threshold determined experimentally as 75%.

Both our margin smoothness prediction method and the approach proposed by [7] use a ribbon (inner and outer) around the nodule boundary and the mean difference in intensity to predict the margin smoothness. However, our method includes additional features in the form of local texture descriptors that capture local intensity variation. These additional features have shown to the improved performance and robustness of the margin smoothness prediction.

Calcification

Calcification in US image is defined as a small and bright fleck in the image reflecting calcium growth on or inside the nodule. Identifying calcifications is known to be challenging due to their variant size, shape and brightness. Certain types of benign characteristics such as the colloids can be easily confused with calcifications in US images. To meet the challenge, we develop a two-stage process where possible candidates for calcification are first detected, and then classified into different classes.

Candidate Detection: For detecting candidates, we adopt the algorithm in [16] that uses a superpixel-based weak detector to propose calcification candidates based on brightness and variance features. Although the algorithm identifies calcification candidates well, it produces many false positive candidates. To overcome this limitation, we propose the following deep learning solution.

Calcification Identification: A Deep Convolutional Neural Network (DCNN) model, known as CaNet, is designed and optimised to validate whether the candidates are actual calcifications. Automatic architecture search often involves training one neural network to optimise the architecture of another neural network. The proposed method performs two consecutive tasks: first, searching for an optimal CNN architecture and hyperparameters using Bayesian Optimization tailored for calcification US images; second, training the CNN model to classify calcification images. For both tasks, 5-fold stratified cross validation was applied with one fold used for architecture optimization and all 5-folds for modelling and evaluating the optimal architecture.

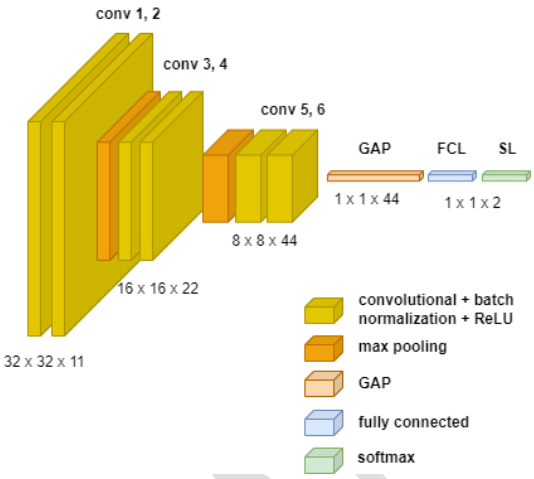


Fig. 4 Optimal CNN architecture for calcification classification.

The initial backbone architecture of CaNet consists of an Input Layer (IL), Convolution Block (CB), Max-Pooling Layer (PL) with stride 2×2, Average Pooling Layer (GAP), Fully Connected Layer (FCL), Softmax Layer (SL), and Classification Output Layer (COL) for two classes (calcification or none-calcification). The CB consists of three layers in the following order: 3×3 Convolutional Layer (CL) with stride 1, Batch Normalization Layer (BNL), and Relu Layer (RL). The IL is set to the size of $32\times32\times1$ to accommodate the small size of the calcification ROI proposed by the weak detector. The hyperparameters were carefully set as follows: initial learning rate to 10^{-4}; optimizer as stochastic gradient descent with momentum; epoch number to 4000; and batch Size to 128. With the architecture defined, the number of CB, the structure of the CB (i.e. the number of filters in the CL) and the type of the PL were determined by the Bayesian Optimization (BO) algorithm [18]. The objective function is defined as the classification error rate. A surrogate model is constructed using the Gaussian Process model, and expected improvement is used as the acquisition function. 30 iterations were performed to search for the optimal parameters. To reduce the likelihood of model overfitting, we also used the BO algorithm to search for optimal L2 regularization value between 10^{-10} to 10^{-2}.

CaNet architecture and model were optimized and trained on a specifically collected dataset of 405 images, where the locations of all calcifications are pinpointed by a radiologist with 15 years of experience. Calcifications in the training set were augmented using mirroring and Singular Value Decomposition (SVD) method [17] with 3 compression ratios (25%, 35% and 45%), finally resulting in 888 calcification candidates and 1723 none-calcification candidates at a ratio of roughly 1:2. The first fold was used to find the optimal CaNet network and Fig 4 shows the details of the optimal architecture with the optimised L2 value of 5.1540e-4 . For the second task, CaNet achieved 81.5% overall accuracy over 5-fold cross-validation, 89.1% specificity (no calcification) and 80% sensitivity (calcification). CaNet model with the highest combined sensitivity and specificity was selected and used in the later stages of identifying micro and macro calcification.

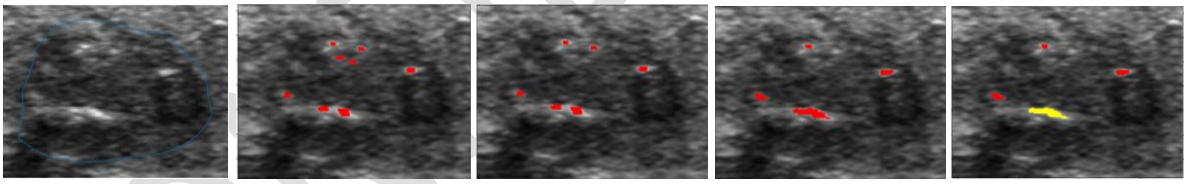
Micro/Macro Calcification Identification: Using the CaNet model, we obtain a set of confirmed candidates. However, it is important to highlight that many confirmed candidates are very small in size due to the nature of the weak detector used. These small candidates represent macrocalcification only partially rather than its entirety.

So, a region growing method is applied to restore the candidates to their appropriate sizes and shapes. In particular, the region growing method uses an iterative flood flow algorithm in comparing the mean brightness of the grown region with its eight neighbours, using the highest brightness value within the candidate as the seed and expanding the region by including new neighbours until they differ significantly from the mean calculated. This naïve region growing method may easily suffer from contrast variations in ultrasound images, occasionally resulting in region overgrowth. Therefore, the growing is counterbalanced using the Speeded-Up Robust Features (SURF) [19]. In our implementation, we have limited the growing region within the areas of 10 strongest SURF descriptors detected, which not only prevents overgrowing calcification regions, but also helps reducing false-positive calcifications. After each detected candidate has been restored to its appropriate size and shape, we extract several features for discriminating micro and macro calcifications. Nodule size, which can be measured by the pixel areas S_c, is an obvious descriptor to distinguish micro from macro calcifications. Some macrocalcifications appear in a line or pseudo-linear shape, which can be captured using circularity o_C. Finally, macrocalcifications often cast acoustic shadows below them. To capture such shadows, we crop the areas immediately above and below the candidate and use the difference between the average brightness of the two areas as the shadow feature Δ_C.

Finally, the grown candidates are classified into micro or macro calcification using the rule in Eq. 6:

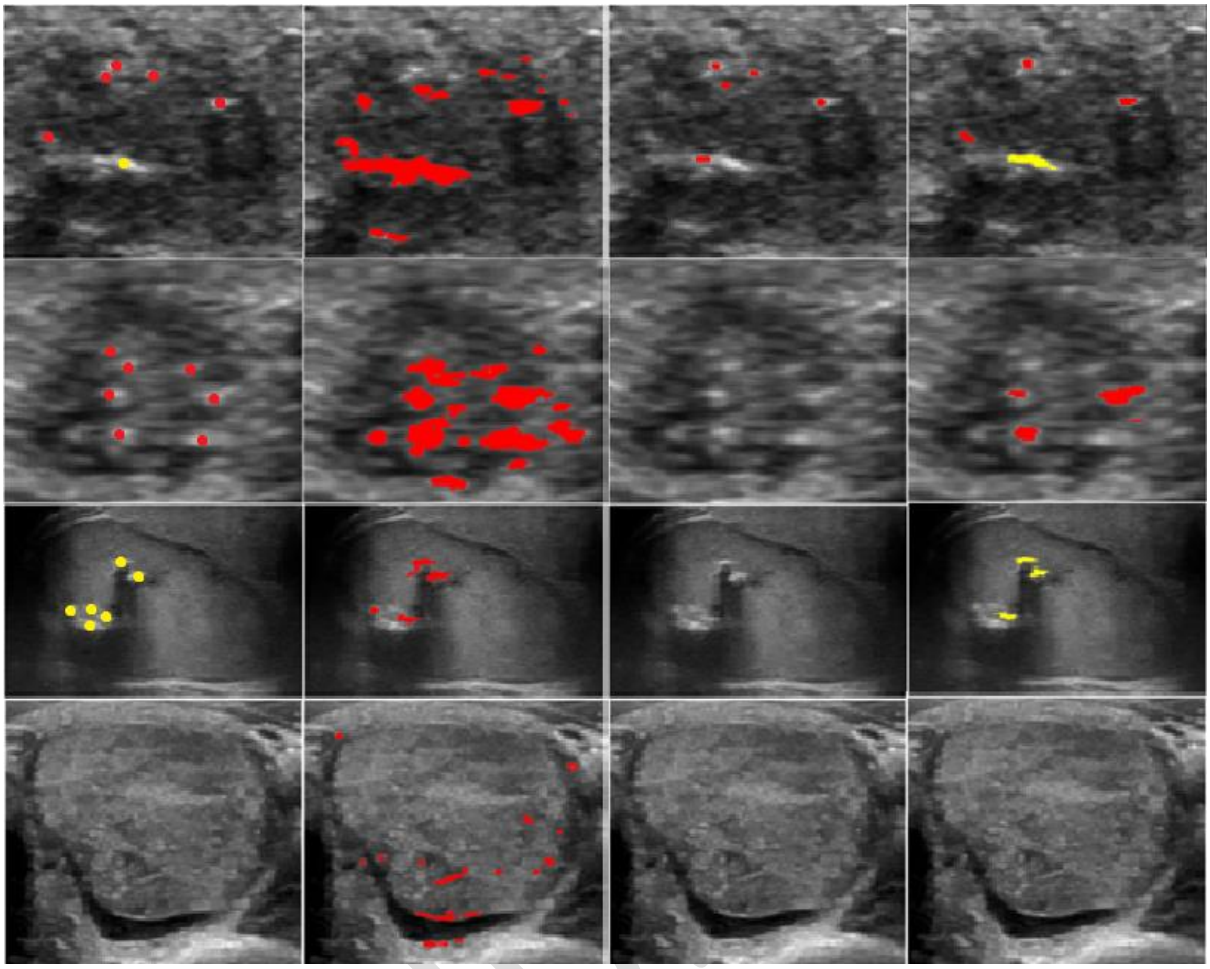
$$\begin{cases} \text{Macro}, & \text{if } S_C > t_{A1} \\ \text{Macro}, & \text{if } S_C > t_{A2} \wedge o_C > t_o \\ \text{Macro}, & \text{if } \Delta_C > t_\Delta \\ \text{Micro}, & \text{else} \end{cases} \tag{6}$$

where t_{A1}, t_{A2}, t_{Cir} and t_Δ are four thresholds for identifying macrocalcifications, which are empirically decided as 200, 95, 0.78 and 50 respectively. Fig. 5I demonstrated the effect of each stage of the proposed calcification detection method when analysing a thyroid nodule with both micro and macro calcifications.



(a) (b) (c) (d) (e)

Fig. 5 I Illustration of Calcification Detection. (a) ROI image; (b) calcification candidates proposed by the weak detector; (c) calcification candidates validated by the CaNet; (d) calcification candidates after growing; (e) classification outcome; red: microcalcification, yellow: macro calcification.



(a) (b) (c) (d)

Fig. 5 II Comparison between Different Calcification Detection Methods. (a) ROI image with calcifications pinpointed by experienced radiologist; (b) top-hat based method [20]; (c) superpixel based method [16]; (d) our proposed method; red: microcalcification, yellow: macrocalcification.

Comparing to the existing methods in the literature, our proposed method strikes a balance between the detected two types of calcification, false alarms and missed cases through the three stage detection process. As shown by examples in Fig.5 II, the morphology-based method [20] is over sensitive, severely suffering from false positive detections whereas the superpixel-based method [16] tends to be under sensitive, failing to detect calcifications in some images.

4. Experiment Results

This section presents the experiment results for evaluating the effectiveness of the various methods proposed in Section 3. All experiments were conducted on an Intel Xeon workstation with CPU@2.90GHz, 16GB RAM, NVIDIA RTX A2000 GPU, and running MATLAB R2020b 64-bits version. With the three patches of images, we used the training patch as the main reference for developing the proposed methods. We then use the validation patch to validate the robustness of the proposed methods and fine-tune the algorithm parameters and the relevant thresholds. For the special characteristic that involves training classification models, such as calcification, to fully utilize the data, we merged the training and validation patches in a 10-fold cross-validation process for model training, evaluation and then selection.

Our experiment consists of two tests. In Test 1, we evaluated the proposed methods against the labels given by one radiologist with 15 years of experience. In Test 2, we evaluated the methods against the labels given by the first radiologist and then confirmed and agreed by another radiologist with similar years of experience. The test results are presented in Table 1.

Table 1 Performance Summary of the Propose Methods in Tests 1 and 2.

Nodule Characteristic Descriptors	**Descriptor Subtypes**		**Test 1**				**Test 2**			
			No. of Cases		**Test Accuracy**		**No. of Agreed Cases**		**Test Accuracy**	
Margin Irregularity	Irregular		96		93.8%		89		94.4%	
	Regular		61		88.5%		43		90.7%	
	Overall (2 Classes)		***157***		***91.7%***		***132***		***93.2%***	
Margin Smoothness	Not Clear		139		89.9%		125		90.4%	
	Clear		18		61.1%		13		77.2%	
	Overall (2 Classes)		***157***		***86.6%***		***138***		***89.1%***	
Calcification	No Calcification		83		94.0%		74		98.7%	
	Calcification	Micro	74	61	83.8%	70.5%	66	42	89.4%	71.4%
		Macro		13		69.2%		9		77.8%
	Overall (2|3 Classes)		***157***	***157***	***89.2%***	***82.8%***	***140***	***125***	***94.3%***	***88.0%***

The table shows our algorithms achieved overall accuracy well above 80% for all three nodule characteristics. The methods perform better on the agreed cases by multiple radiologists than cases labelled by a single radiologist. In general, the algorithms perform better on characteristics that are clearly defined than those where there is more room for different interpretations. For instance, high levels of accuracy are achieved for Margin Irregularity whereas the algorithms' performances on margin smoothness and calcification are relatively lower.

At subtype level, the algorithm performances vary substantially due to uneven distributions of the subtypes particularly for those characteristics with a greater degree of subjectivity. The difficulty faced by the algorithm development is which radiologist's labels should be based on as the ground truth. The difficulty is more severe when the number of cases of a subtype is small. Margin characteristics are also known for their subjective nature, where we found that radiologists agree more on irregular (89 of 96, 92.71%) and unclear cases (125 of 139, 89.93%) than on regular (43 of 61, 70.49%) and clear cases (13 of 18, 72.22%). This is because boundaries of malignant nodules tend to have more distinctive appearances than those of benign nodules, and hence radiologists often have differences when classifying boundaries of "benign-looking" nodules. Despite these subjective factors, the proposed methods still achieved good performance on the margin characteristics.

The test results on calcification show that our calcification method achieved good overall accuracy, but better performance on none-calcification than calcification at subtype level. This performance bias is understandable as none-calcification cases appear more frequently in clinics. The results also show that inter-observer variability is quite substantial for calcification; the radiologists agree more in calcification and none-calcification (140 of 157, 89.17%) than in micro and macro calcifications (51 of 74, 68.92%). It is worth noting that radiologists often use the measurement scale marked on the side of US image as an aid when classifying micro and macrocalcifications whereas the algorithms have not made such a reference.

We have also compared our proposed CNN model against other novel CNN models adapted for calcification detection. In particular, we tuned two poweful CNNs, VggNet19 and ResNet101, using transfer learning approach for calcification image classification. The architectures of VggNet19 and ResNet101 were adapted by replacing and fine-tuning the last fully connected layer and the softmax layer of each network. The last fully connected layer was also replaced by a new fully connected layer for two classes (calcification, no calcification). For a systematic and fair comparison, we set the network parameters for both models as follows: 20 epochs, initial learn rate = 0.0001, and mini-batch size = 4. The other parameters were set as default values of each networks. The results in Table 2 show that our proposed CaNet model achieved better specificity. Although CaNet model has lower sensitivity than the other two models, it is worth mentioning that our CaNet achieved less biased results when classifying micro and macro calcifications. It is also worth noting that both transfer learning models had a required input size of 224×224x3, which did not fit most of the calcification candidates due to their small sizes. To resolve the issue, the candidates were resized using bicubic interpolation. However, it is a known fact that up-samplings may easily cause overfitting and fuzziness in the model trained. We believe that this explaines why our proposed model has achieved better and less biased results as it fits better to the small input size.

Table 2: Comparison of Using Different CNN Models for Calcification Detection.

Calcification Labels		**VGGNet-19**		**ResNet-101**		**CaNet**	
No Calcification		83.1%		80.7%		94.0%	
Calcification	Micro	85.1%	75.4%	85.1%	70.5%	83.8%	70.5%
	Macro		53.8%		61.5%		69.2%
Overall		84.1%	77.7%	82.8%	75.2%	89.2%	82.8%

5. Discussions

Our tests in the experiment section have shown promising results from our proposed method. In this section, we will discuss several issues concerning optimization of our models and algorithms, including the optimised thresholds, other TIRADS characteristics and constructing robust models.

Threshold Tuning

The proposed methods for margin and calcification use several thresholds. To determine the best configurations empirically, we conducted a gradient descent-based search on the validation set. The search aims to maximise the overall validation accuracy while maintaining a balance between the sub-class accuracies. Fig 6 – 8 illustrate how threshold setting may affect modelling performance.

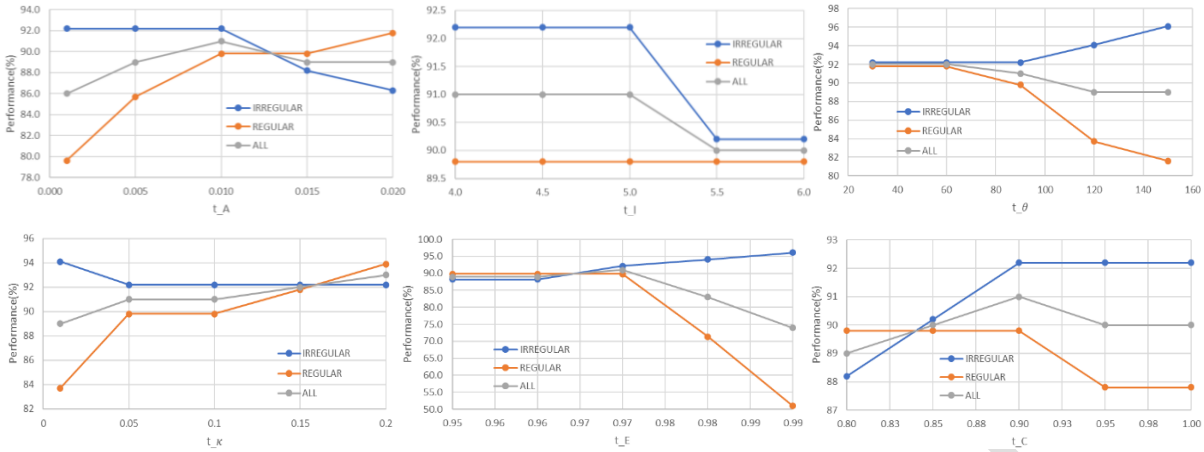


Fig. 6 Performance of different thresholds used in margin irregularity analysis.

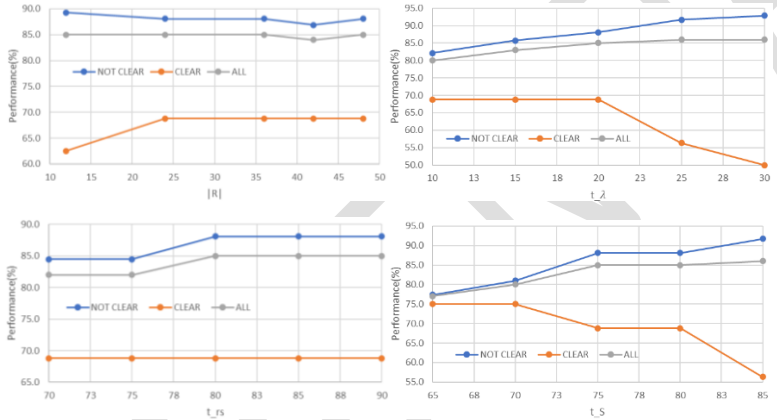


Fig. 7 Performance of different thresholds used in margin smoothness analysis

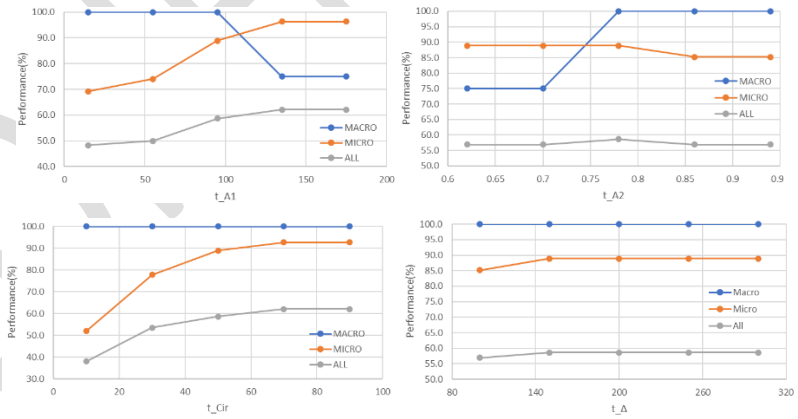


Fig. 8 Performance of different thresholds used in calcification analysis.

We also found that the size of a nodule in an image may affect radiologist's decisions when classifying margin characteristics. Lobulations in large nodules for example may appear less severe than the same kind of lobulations in smaller nodules. Therefore, we have further altered some thresholds according to the size of the nodule for better robustness. In particular for margin irregularity, we set t_A = 0.01, t_l = 2, t_κ= 0.2 for nodules with a minimum resolution less than 50 pixels, and set t_κ= 0.15 for nodules with a minimum resolution between 50 and 100 pixels. On contrast, margin smoothness is less affected by nodule size because the characteristic relates to

textures around the margin. Most features used in the proposed methods expect larger nodules for statistical reliability of the extracted information. In reality, there are nodules of very small sizes. Therefore, instead of using the thresholds universally defined, we set $|R|$ = 12, $t_\lambda = 15$, and t_S = 80 for nodules with a minimum resolution that less than 100 pixels, and $|R|$ = 24 for nodules with a minimum resolution between 100 and 250 pixels.

Table 3 Performance Summary of Test 1 based on Cancer Type and Nodule Size.

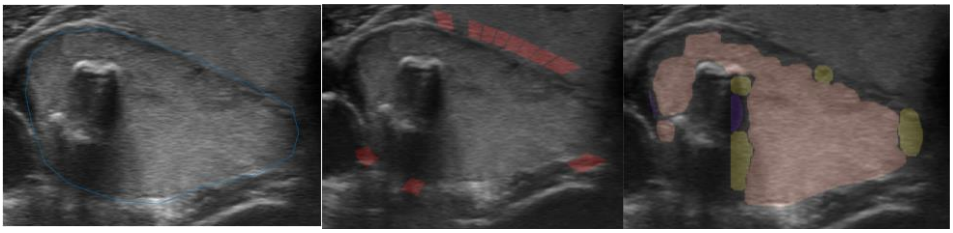
Test 1	Benign	Malignant	Size<150 pxl	Size≥150 pxl
Margin Irregularity	86.0%	93.9%	92.2%	91.0%
Margin Smoothness	81.4%	88.6%	88.9%	83.6%
Calcification	69.8%	87.7%	83.3%	82.1%

As presented in table 3, we found that the proposed methods tend to perform better for malignant nodules than benign ones. The performance bias may be partially caused by the unbalanced training dataset, which can be improved by enrolling or augmenting more benign cases. We also have relatively poor accuracy for calcification on benign nodules. It is worth noting that calcification is often associated with malignancy and rarely appears in benign nodules. Also, fibrosis in benign tumours can be easily confused with microcalcifications. Further research is needed to better understand such rare and confusing cases for developing better solutions. Additionally, the experiment results also show that the margin irregularity measure performs well and is robust across different nodule sizes because our method has considered both global and local margin irregularities. Margin smoothness also performs well but is slightly better towards small nodules. We believe the performance deterioration for large nodules is due to the excessive space covered by each region, indicating that it may be appropriate to increase the value of $|R|$ when analysing large nodules.

Shape and Echogenicity Analysis

Besides margin and calcification, the TIRADS guidelines also define other US characteristics such as shape and echogenicity. Shape describes the orientation of the nodule growing. To provide a complete automated solution, we proposed a simple shape classification algorithm consisting of three steps. First, a polygon shape (or bounding box) is constructed based on the set of coordinates on the ROI boundary. An exhaustive search is then conducted horizontally and vertically within the polygon to locate the maximum width w_{max} and the maximum height h_{max}. The nodule is then classified as "taller-than-wide" if $h_{max} \geq w_{max}$; otherwise, "wider-than-tall". Using the test set images, this simple algorithm achieved high accuracy of 98% over cases labelled by the single radiologist and 100% for the agreed cases by two radiologists.

Another important US characteristic is echogenicity which is reflected by the pixel intensity values in the nodule region of the image. We proposed a simple algorithm to identify the echogenicity type by comparing the intranodular intensity with that of the surrounding areas. The algorithm first divides the areas around the nodule into sub-bands (or small regions) and studies the mean and variance on their intensities. Sub-bands that being over dark/bright/inconsistent are consider as non-gland area and being excluded from consideration. The median of the remaining ones is then used as the isoechoic reference and the other echogenicity types are determined accordingly using a set of dependent thresholds. At the end, we compare the percentage of the echogenicities contained and chose the most dominant one as the final echogenicity class. Fig. 9 shows some example echos within a nodule.



(a) (b) (c)

Fig. 9 Illustration of Echogenicity Detection: (a) Original ROI; (b) valid reference regions detected, marked in red; (c) echogenicity classification result (purple: very-hypoechoic; yellow: hypoechoic; pink: isoechoic).

The test results also show that our proposed algorithm achieved an overall accuracy of 87.7% against one radiologist labels, and a higher overall accuracy of 90.6% when it was tested against the agreed cases by two radiologists. However, since some rare subtypes such as hyperechoic and very-hypoechoic are extremely difficult to obtain from clinical practice, the test set was too small in the current data collection. Further evaluations are needed to test the reliability of our proposed methods for echogenicity.

Ablation Study

We used 100 randomly selected images and conducted two small scale ablation analyses on margin-smoothness and margin-irregularity to evaluate the contribution of individual features and their combinations to the overall performances of the methods. The test results are shown in Tables 4 and 5.

Table 4: Ablation Study for Margin Smoothness.

Labels for Smoothness	Region Analysis Only	Signal Analysis Only	Region & Signal Features
Clear	37.5%	62.5%	75.0%
Not Clear	97.6%	82.1%	86.9%
Overall	88.0%	79.0%	85.0%

Table 5: Ablation Study for Margin Irregularity.

Irregularity Labels	Stage One		Stage One and Stage Two		
	Convex Variance	Elliptic Variance	Conv. Var + Local	Ellip. Var. + Local	All Features
Irregular	15.7%	96.1%	88.2%	96.1%	92.2%
Regular	100.0%	49.0%	89.8%	46.9%	89.8%
Overall	57%.	73%	89%	72%	91%

The margin-irregularity analysis revealed that between the stage one features elliptic variance (73%) contributes more to the prediction than convex variance (57%). The result also indicates the bias of the features towards different subclasses. However, adding the local features (lobulation and angulation), convex variance performed 17% better than elliptic variance with local features. The evaluation of the combined feature showed an improvement in the performance, with a performance increase of 2% compared to the highest-performing individual feature. Therefore, the features complement each other and enhance the method's performances.

The margin-smoothness results showed that Signal Analysis had a higher contribution compared to Region Analysis. The analysis also demonstrated that the combination of Region and Intensity improved the model's robustness compared to the individual features, suggesting that the two features complement each other.

To analyse the essence of each step of our proposed three-stage calcification detection method, we have performed an ablation study using the 157 images from test 1. Test results showed that the weak detector, CaNet and SURF filter had a 16% impact on average when identifying calcifications. The weak detector was mostly affecting calcifications detections. In comparison, both CaNet and SURF filters were mostly improving false detections. The region-growing method did not contribute much when identifying calcifications but improved the macro calcification classifications significantly (see Table 6).

Table 6: Ablation Study for Calcification.

Calcification Labels		**Weak Detector***		**CaNet**		**SURF Filter**		**Region Growing**		**All**	
No Calcification		89.2%		51.8%		53.0%		94.0%		94.0%	
Calcification	Micro	51.4%	44.3%	97.3%	93.4%	97.3%	93.4%	82.4%	75.4%	83.8%	70.5%
	Macro		0.0%		23.1%		23.1%		15.4%		69.2%
Overall		71.3%	65.0%	73.3%	65.6%	73.9%	66.2%	88.5%	80.3%	89.2%	82.8%

*****The weak detector was replaced by the top-hat detector for the ablation study

Margin Smoothness Sensitivity Analysis to the Precision of the Region of Interest (ROI)

Whilst we are unable to apply other margin smoothness methods to our dataset due to the differing objectives between the studies as mentioned in Section 2, we have conducted an analysis about the sensitivity of our method when the delineated RoI does not precisely align with the lesion boundary. We purposely introduced various degrees of misalignment by applying random shifts to the initial RoI. The process begins with determining the ribbon width for each lesion (refer to Margin Smoothness section for the ribbon's definition) with a defined shift range of ±20%. For each lesion, a random shift value is chosen from the predetermined range. This process is repeated ten times on the 100 randomly selected images, and the performances are presented in Table 7.

Table 7: Sensitivity Analysis of Margin Smoothness method to the precision of the delineated region of interest (ROI) i.e., the performances when the region of interest does not precisely align with the boundary of the lesion.

Iteration	**0**	1	2	3	4	5	6	7	8	9	10
Clear	**75.00**	56.25	56.25	56.25	68.75	56.25	56.25	68.75	50.00	62.50	68.75
Not Clear	**86.90**	94.05	90.48	89.29	94.05	90.48	90.48	92.86	91.67	95.24	89.29
Overall	**85.00**	88.00	85.00	84.00	90.00	85.00	85.00	89.00	85.00	90.00	86.00

Table 7 highlights 10-iteration performances of the ROI precision analysis. Iteration 0 shows the performance with the original delineated ROI, while iterations 1-10 show the performances after applying the random shifts to the ROI. The overall performance indicates the method is not over sensitive to variations in ROI precision i.e., the performance is relatively stable even with imprecise ROI. The performance has a standard deviation of 2.25. This conclusion is consistent with "Not Clear" performance which also has a standard deviation of 2.50. However, the "Clear" performances indicate high variability with a standard deviation of 7.82. This variability is attributed to the smaller sample size of "Clear" cases, where any misclassification will significantly impact the performance.

6. Conclusion

In this paper, we presented several methods for detecting US image characteristics of thyroid nodule for margin irregularity, margin smoothness, and calcification. The proposed method for margin classification have exploited new geometrical and texture features effectively. Our novel three-stage approach for calcification identification utilizes super-pixels and a convolutional neural network optimized for this purpose. Finally, a simple method for nodule shape and an initial algorithm for echogenicity using the thyroid gland as the main reference have been described. Our methods have shown good performances in identifying the US image characteristics of thyroid nodules with overall accuracies from 82.8% to 98.1% when tested on US images collected from two hospitals and labelled by multiple experienced radiologists. Encouraged by the results, we will continue improving our algorithms for thyroid characteristics analysis and expand our work to estimating TIRADS scores for the nodule and identify level of malignancy. Furthermore, we plan to adapt the methods for identifying characteristics for other kinds of lesions such as breast lesions and lymphoma. Finally, we will compare the performance accuracies of our methods with nodule contour extracted from automatic segmentation.

Acknowledgement

This research is sponsored by TenD.AI Medical Technology.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA and Jemal A (2018): "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", CA Cancer J Clin, 68(6), pp.394-424.

2. Kwak JY, Han KH, Yoon JH, Moon HJ, Son EJ, Park SH, Jung HK, Choi JS, Kim BM., and Kim EK (2011): "Thyroid imaging reporting and data system for US features of nodules: a step in establishing better stratification of cancer risk", Radiology. 260(3), pp.892 - 899.

3. Tessler FN, Middleton WD, Grant EG., Hoang JK., Berland LL, Teefey SA, Cronan JJ, Beland MD, Desser TS, Frates MC, Hammers LW, Hamper UM, Langer JE, Reading CC, Scoutt LM, and Stavros AT (2017): "ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee", Journal of the American College of Radiology. 14(5), pp.587-595.

4. Wang H, Yang Y, Peng B, and Chen Q (2017): "A thyroid nodule classification method based on TI-RADS", Proc. SPIE 10420, Ninth International Conference on Digital Image Processing (ICDIP 2017), 1042041.

5. Wu MH, Chen CN, Chen KY, Ho MC, Tai HC, Wang YH, Chen A, and Chang KJ (2016): "Quantitative analysis of echogenicity for patients with thyroid nodules", Scientific reports, 6, 35632.

6. Zulfanahri NHA, Nugroho A, Frannita EL and Ardiyanto I (2017): "Classification of thyroid ultrasound images based on shape features analysis", 10th Biomedical Engineering International Conference (BMEiCON), pp. 1-5.

7. Zhuang Y, Li C, Hua Z, Chen K, and Lin JL (2018): "A novel TIRADS of US classification", Biomedical Engineering Online, 17, 82.

8. Long J, Shelhamer E, and Darrell T (2015): "Fully convolutional networks for semantic segmentation", In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), Boston, MA, USA. pp. 3431–3440.

9. Nugroho HA, Frannita El, and Hutami AHT (2020): “Thyroid nodules stratification based on orientation characteristics using machine learning approach”, 3rd International Conference on Computer and Informatics Engineering (IC2IE), pp. 52-57.

10. Nugroho, HA, Frannita EL, Nugroho A, Zulfanahri AI and Choridah L (2017): “Classification of thyroid nodules based on analysis of margin characteristic”, 2017 International Conference on Computer, Control, Informatics and its Applications (IC3INA), pp. 47-51.

11. Ha EJ, Na DG. and Baek JH (2021): “Korean thyroid imaging reporting and data system: current status, challenges, and future perspectives”, Korean Journal of Radiology, 22(9), pp. 1569-1578.

12. Smith D and Botz B (2021): “European thyroid association TIRADS”, Reference article, Radiopaedia.org. (accessed on 02 May 2022) https://doi.org/10.53347/rID-68341

13. Zhou J, Yin L, Wei X, Zhang S, Song Y, Luo B, Li J, Qian L, Cui L, Chen W, Wen C, Peng Y, Chen Q, Lu M, Chen M, Wu R, Zhou W, Xue E, Li Y, Yang L, Mi C, Zhang R, Wu G, Du G, Huang D, and Zhan W (2020): “Superficial organ and vascular ultrasound group of the society of ultrasound in medicine of the Chinese medical association”, Chinese Artificial Intelligence Alliance for Thyroid and Breast Ultrasound, Nov;70(2):256-279.

14. Qi Q, Zhou A, Guo S, Huang X, Chen S, Li Y and Xu P (2021): “Explore the diagnostic efficiency of Chinese thyroid imaging reporting and data systems by comparing with the other four systems (ACR TI-RADS, Kwak-TIRADS, KSThR-TIRADS, and EU-TIRADS): A Single-Center Study”, Front Endocrinol. 12:763897.

15. Horvath E, Majlis S, Rossi R, Franco C, Niedmann JP, Castro A and Dominguez M (2009): “An ultrasonogram reporting system for thyroid nodules stratifying cancer risk for clinical management”, Journal of Clinical Endocrinology and Metabolism, 94(5), pp. 1748-1751

16. Ren L, Liu Y, Tong Y, Cao X, and Wu Y (2020): “Calcification segmentation based on a different scales superpixels saliency detection algorithm”, Journal of Ultrasound in Medicine & Biology, 46(12), pp. 3404-3412

17. Zhu Y, AlZoubi A, Jassim S, Jiang Q, Zhang Y, Wang Y, Ye X, and Du H (2021): “A generic deep learning framework to classify thyroid and breast lesions in ultrasound images”, Journal of Ultrasonics, 110, 106300

18. Radhakrishnan, R. and AlZoubi, A. (2022). Vehicle Pair Activity Classification using QTC and Long Short Term Memory Neural Network. In Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP, ISBN 978-989-758-555-5; ISSN 2184-4321, pages 236-247. DOI:10.5220/0010903500003124

19. Bay H, Ess A, Tuytelaars T, and Gool VL (2008): “Speeded-up robust features (SURF)”, Journal of Computer Vision and Image Understanding, 110(3), pp. 346-359

20. Dong Y, Gao X and Wang Y(2006): “A Top-hat Based Calcifications Detection Method in Mammograms”, Journal of Image and Graphics, 11(12), pp. 1839-1843