Title Page

A Generic Deep Learning Framework to Classify Thyroid and Breast Lesions in Ultrasound Images

Yi-Cheng Zhu[1*], Alaa AlZoubi[2*], Sabah Jassim[2], Quan Jiang[1], Yuan Zhang[1], Yong-Bing Wang[3], Xian-De Ye[3], Hongbo DU [2#]

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

[2] School of Computing, University of Buckingham, Buckingham, UK

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

* equal contribution

# corresponding author: Hongbo Du

School of Computing
University of Buckingham
Hunter Street, Buckingham, MK18 1EG, United Kingdom
Email: hongbo.du@buckingham.ac.uk
Tel: +44 (0)1280 828298 / 828322

Abstract

Breast and thyroid cancers are the two common cancers to affect women worldwide. Ultrasonography (US) is a commonly used non-invasive imaging modality to detect breast and thyroid cancers, but its clinical diagnostic accuracy for these cancers is controversial. Both thyroid and breast cancers share some similar high frequency ultrasound characteristics such as taller-than-wide shape ratio, hypo-echogenicity, and ill-defined margins. This study aims to develop an automatic scheme for classifying thyroid and breast lesions in ultrasound images using deep convolutional neural networks (DCNN). In particular, we propose a generic DCNN architecture with transfer learning and the same architectural parameter settings to train models for thyroid and breast cancers (TNet and BNet) respectively, and test the viability of such a generic approach with ultrasound images collected from clinical practices. In addition, the potentials of the thyroid model in learning the common features and its performance of classifying both breast and thyroid lesions are investigated. A retrospective dataset of 719 thyroid and 672 breast images captured from US machines of different makes between October 2016 and December 2018 is used in this study. Test results show that both TNet and BNet built on the same DCNN architecture have achieved good classification results (86.5% average accuracy for TNet and 89% for BNet). Furthermore, we used TNet to classify breast lesions and the model achieves sensitivity of 86.6% and specificity of 87.1%, indicating its capability in learning features commonly shared by thyroid and breast lesions. We further tested the diagnostic performance of the TNet model against that of three radiologists. The area under curve (AUC) for thyroid nodule classification is 0.861 (95% CI: 0.792-0.929) for the TNet model and 0.757-0.854 (95% CI: 0.658-0.934) for the three radiologists. The AUC for breast cancer classification is 0.875 (95% CI: 0.804-0.947) for the TNet model and 0.698-0.777 (95% CI: 0.593-0.872) for the radiologists, indicating the model's potential in classifying both breast and thyroid cancers with a higher level of accuracy than that of radiologists.

Key words: Thyroid Cancer, Breast Cancer, Ultrasonography, Cancer Recognition, Deep Convolutional Neural Network

Abbreviations

US = Ultrasonography, MRI = Magnetic Resonance Imaging, CT = Computed Tomography, CNN = Convolutional Neural Network, ROI = Region of Interest, SVD = Singular Value Decomposition, ROC = Receiver Operating Characteristics

Funding

The work was supported by the TenD Innovations. Sponsors had no role in the study design or performance, data acquisition, and interpretation, or article preparations.

1. Introduction

Breast cancer is the most commonly diagnosed cancer in women, and thyroid cancer is among the top five most common cancers in women globally [1]. Magnetic resonance imaging (MRI), computerized tomography (CT), and ultrasonography (US) have become indispensable imaging modalities that are widely used to screen and aid the diagnosis of breast lesions and thyroid lesions nowadays. Compared with MRI and CT, US is a universally used imaging modality that is non-invasive, non-radiative, and of lower cost. The accuracy of US-based diagnoses of thyroid or breast cancers, however, largely depends on the experience and cognitive capabilities of individual radiologists [2]. Due to such challenges, many studies have reported the usefulness of computer-aid diagnosis (CAD) systems [3]. Exploiting machine learning and computer vision techniques, a CAD system attempts to extract morphological and texture features from ultrasound images and train effective models based on the extracted features to classify the status of malignancy for the thyroid and breast lesions. However, conventional machine learning algorithms designed specifically for extracting morphological features (such as regularity and uniformity of lesion boundaries [4]) or texture features (such as local binary patterns (LBP) [5], grey level co-occurrence matrices (GLCM) [6]) often require “hand-crafted” optimal combinations and complex processes of image pre-processing, feature extraction and classification. The overall performance of such a system is heavily influenced by factors such as image modalities, image qualities, similarity in morphology of lesions, type of cancers, etc., and their capability of discriminating benign and malignant lesions is often limited [7].

Recently, convolutional neural networks have shown their outstanding capabilities in object recognition especially for the largescale visual recognition tasks, their strengths in feature learning (such as color, textures and shape), and their ability to capture discriminative and robust information from images by applying convolution operations with suitable filters over a sequence of convolutional layers [8]. Deep learning has also been introduced into CAD systems to classify US images [9-11] or microscopic images [12] of various types of tumours including thyroid and breast lesions. Existing research mainly focuses on customizing and modifying known CNN architectures specifically chosen for a certain type of cancer. However, none of the published studies of lesion classification have worked on a generic deep learning architecture for building models to classify both thyroid and breast lesions in ultrasound images. Such a generic approach of deep learning solutions simplifies the process of constructing classification models for multiple types of cancer and can be desirable in clinical practice. Previous evidences suggest that the chance of having breast and thyroid cancers in the same female patients is greater than that of the general population [13,14]. A possible association between breast and thyroid cancer has also been demonstrated, including shared hormonal risk factors and genetic susceptibility [15]. Furthermore, thyroid and breast cancers do share common image characteristics under high frequency ultrasound scans such as malignant lesions with a taller-than-wide shape ratio, hypo-echogenicity, and ill-defined margins [16,17]. This observation provides a strong motivation for developing a generic convolutional neural network (CNN) model that can be used to classify breast and thyroid cancers.

The key contributions of this paper include: (1) a generic CNN-based modelling framework suited for both thyroid and breast lesion classification based on a modified version of an known architecture [18], (2) a novel singular value decomposition (SVD) technique for data augmentation to enlarge the training set and generalize the trained models, (3) trained CNN models on thyroid or breast images captured from US machines of different makes that can learn common features of both types of lesions, and (4) an evaluation showing that the trained TNet and BNet perform well and that the TNet model either matches or even outperforms experienced radiologists in classifying both breast and thyroid lesions.

2. Materials and Methods

This section presents the main aspects of the proposed method including data acquisition and annotation, data augmentation and generic CNN modelling.

2.1 Patients and Lesions

This retrospective study was approved by the Ethics Committee of Shanghai Pudong People's Hospital China (referred to as "the Hospital"), who waived the requirement for informed consent, and by the Research and Ethics Committees of University of Buckingham UK. The study consisted of a cohort of 1,611 female patients (66.36 ± 8.67 years of age, range between 43 and 95 years old) from the Hospital between October 2016 and December 2018. After excluding 14 patients because of missing data, 821 patients with thyroid lesions and 776 patients with breast lesions were included (Figure 1). A total of 719 thyroid lesions (298 malignant and 421 benign) and a total of 672 breast lesions (299 malignant and 373 benign) were used to build and validate the classification models (Figure 1). All lesions were confirmed by histopathological assessment of tissue samples obtained via biopsy or surgery.

2.2 US Image Acquisition

All thyroid and breast gray-scale US examinations were performed in the Hospital using US machines of five different makes and models including Siemens Oxana 2, Siemens S3000, Toshiba Apolio 500, GE Logic E9, and Philips Epic 7 with a high-frequency linear probe (5-12 MHz for both thyroid and breast imaging). These machines are most commonly used to capture US images in real clinical practice, and we wanted to ensure that the trained CNN models would be robust. Both longitudinal and transverse planes of the thyroid lesions and breast lesions were obtained. For instance, among the lesions for developing the DCNN models (see Section 3.1), 525 (73.0%) and 248 (36.9%) longitudinal planes of the thyroid lesions and breast lesions were respectively obtained. Lesions with the largest diameter in US were selected for patients with more than one lesion. All images were acquired and stored in RGB format. The TI-RADS [19] and BI-RADS [20] were referred to evaluate the malignancy risk of each lesion stratified by its US patterns composed of the integrated solidity, echogenicity, and suspicious US features of each lesion.

2.3 CNN based Cancer Recognition

2.3.1 US Image Pre-processing

Since the adopted network architecture [18] was pre-trained on images with a single object occupying the entire scene, to satisfy the training requirements, the acquired US images were subjected to preprocessing. The region of interest (RoI), i.e. the lesion area of the image, was cropped from the whole ultrasound image for accurate recognition. A free-hand cropping software tool was developed using MATLAB. The tool enables radiologists to identify pixel points marking the border of a lesion, and the tool collects the coordinates of the points. Using the software tool, all RoIs were first cropped manually by a radiologist with at least 5 years of experience in both thyroid and breast US (Figure 2) and then checked by a senior radiologist with >15 years of experience in thyroid and breast imaging. A rectangular bounding box was generated for each lesion by fitting the border points into minimum-area-rectangle. The image within the bounding box is known as an RoI image herein. RoI images of lesions were then used as input images for CNN model training and testing.

2.3.2 Data Augmentation

Training and tuning an architecturally complex DCNN of a large size, such as VGGNet [18], requires a large number of training images. Large datasets comprising thousands of ultrasound images annotated with accurate class labels (i.e. the ground-truth) are always challenging and difficult to obtain and thus are in short supply. One possible way to overcome this issue and reduce potential model overfitting is to artificially enlarge the training set available using label-preserving transformations, known as data augmentation [21]. In this study, we proposed two types of techniques to augment the cropped US RoI images: Geometric methods and Singular Value Decomposition method.

2.3.2.1 Geometric Methods

Rotation and mirroring alter image geometry of the image by mapping the individual pixel values to new destinations. Here, both methods change the original RoI image to a new position and orientation while preserving the shape of the class representation within the image. For rotation, each RoI image was rotated counterclockwise around the center of the RoI with degrees of 90, 180, and 270. For mirroring, a reflected duplication of an RoI image was generated by flipping the image across its vertical axis. These geometric methods generated four artificial images from each RoI image. Image features such as textures, echogenicity, margin characteristics are not affected by the operations. Both methods were considered to be computationally efficient as they were applied directly on the image matrix.

2.3.2.2 Singular Value Decomposition (SVD)

An image compression-related SVD-based scheme was used to generate approximate images with different degrees of compressed contents while preserving the geometric features of the original RoI image. The images were obtained by ranking the information content according to the levels of its importance in the original image data. In other words, we use SVD method to disclose the structure of the image matrix to obtain the further compression of the original RoI images. The working principle of the method is explained as follows.

A cropped RoI image of r rows and c columns of pixels in the RGB color space forms three $r \times c$ matrices $M\{R, G, B\}$ respectively representing the RGB channels. The singular value decomposition for each of the three matrices is a factorization of the form:

$$M_{\{R,G,B\}} = U \Sigma V^T$$

where U is of size $r \times r$, Σ is of size $r \times c$, and V^T is of size $c \times c$. U and V are orthogonal matrices, and Σ is a diagonal matrix whose entries arranged in descending order along the main diagonal. The matrix Σ represents the singular values of M and determines the rank of the original matrix.

The three RGB channels were processed individually and then later stacked back on top of each other to create a new RGB image. For each RoI image, three images were generated with 45%, 35% and 25% ratios of the selected top singular values.

2.3.3 Building CNN Models

The parameters of the CNN model VGG-19 [18] were pre-trained on the ImageNet dataset [8] for the task of object recognition from the images. The network has 47 layers, comprising 16 convolutional and 3 fully connected learnable weight layers. Each convolution layer consists of filter size 3x3 and different number of kernels. The model contains approximately 144 million weight parameters, and the convolutional layers extracts local features such as lines, shapes, edges, and textures that could be transferred for similar visual recognition tasks, such as cancer recognition in ultrasound images.

The layers trained using the CNN [18] and the ImageNet dataset [8] were adapted for cancer recognition. The architecture of the CNN model [18] was adapted by replacing and fine-tuning the last fully connected layer (fc8), the softmax (prob) layer and the output layer (output). Since the images of each cancer type (thyroid and breast) is labelled by either of two classes, a new fully connected layer (fc8') was added for the two classes (indicating benign and malignant). A softmax layer (prob') and a classification output layer (output'), where the output of the last fully-connected layer was fed to a 2-way softmax layer (or normalized exponential function), produce a distribution over the two class labels. In addition, we set the last 'Dropout' layer to 25%. The adaptions result in a generic DCNN architecture which was then used to build the TNet and BNet models for the thyroid and breast cancers respectively. Figure 3 illustrates the modified CNN architecture. The TNet model was trained on thyroid RoI images and the BNet model was trained on breast RoI images.

Training and testing procedures were developed based on the ultrasound RoI images. As an additional preprocessing step, each RoI image was rescaled to 224 x 224 x 3 by using the bicubic interpolation method, augmented using the SVD and the geometric methods, and then fed as inputs to the data layer (data) of the network. The rescaling of RoI images to the target size is to meet the data layer requirement of the adapted CNN architecture [18]. The network hyperparameters were set as follows: iteration number = 9080, initial learn rate = 0.0001, and mini batch size = 8. These configurations were finalized empirically to ensure that the parameters were finetuned for the cancer recognition task. We observed that the model stopped learning after 20 epochs which represents ~9080 iterations. Several different learning rates (0.01, 0.001, and 0.0001) were attempted, and 0.0001gives the best loss without sacrificing speed of training. The other network parameters were set to their default values [18]. Data augmentation, 25% drop out of the last 'Dropout' layer and imbalanced data methods were techniques used to reduce the effect of model overfitting. We found experimentally that using relatively more images of benign cases in the training set reduces the model sensitivity and helps reducing the model overfitting overall.

All experiments were run on an Intel Core i7 desktop, two GPU GeForce RTX™ 2080, CPU@2.30GHz (two processors) with 64.0 GB RAM.

2.4 Observer Study by Radiologists

The test ultrasound images were presented on a standard reporting workstation in random order to three radiologists with 3 to 15 years of experience in both thyroid and breast imaging between them. These radiologists classified each lesion as being either malignant or benign. The clinical information of each patient was withheld from the invited radiologists.

2.5 Statistical analysis

Receiver operating characteristics (ROC) curves were used to demonstrate and compare the diagnostic performance of our deep learning models with that of the experienced radiologists in classifying benign and malignant cases in thyroid cancer and breast cancer. The individual and average sensitivity, specificity and accuracy rate of the three radiologists was used when comparing diagnostic performance. The SPSS (version 25.0, SPSS Inc., Chicago, IL, USA) software was utilized for data analysis. P values <0.05 were considered as statistically significant.

3. Results

3.1 Study population

A total of 672 patients (58.4 ± 16.3 years old) with 672 breast ultrasound images (benign: 373, malignant: 299) (Table 1) and 719 patients (55.3 ± 12.6 years old) with 719 thyroid ultrasound images (benign: 421, malignant: 298) (Table 2) were used in developing (i.e. training and testing) the TNet and BNet models. Two additional sets

(102 thyroid lesions and 104 breast lesions) were set aside for comparing the models against radiologists, where 45 out of 102 thyroid nodules (Table 2) were malignant and 52 out of 104 breast nodules were malignant (Table 1).

3.2 Evaluation of the CNN models

We first performed comparative experiments in order to evaluate the effectiveness of our method, using two different US image datasets (breast and thyroid datasets). First, we used 719 US thyroid images (298 malignant and 421 benign) to evaluate the performance of the TNet model. To determine the classification accuracy, we used 10-fold stratified cross validation. On each iteration, we split the US images into training and testing sets at ratio of 90% to 10% for each class. Among the training examples for each fold, 10% of them were used as validation examples. The TNet model achieved an average accuracy of 86.5% (std = 2.8%), an average true positive rate (TPR) of 83.9% (std = 3.9%) and an average true negative rate (TNR) of 88.6% (std = 4.6%) in classifying thyroid lesions (Table 3). To evaluate the performance of our generic CNN models (TNet), we also used the TNet to classify all breast cases (672 images). The TNet model achieved an average accuracy of 86.6% on classifying breast malignant cases (sensitivity) and 87.1% on classifying breast benign cases (specificity).

We conducted similar classification experiments using the breast US image dataset. This comprised 373 benign images and 299 malignant images. We also used 10-fold cross validation to evaluate the classification accuracy. On each iteration, we split the US images into training and testing sets at ratio of 90% to 10% for each class. The same arrangement for the validation examples as for the TNet was also applied. The BNet model achieved an average accuracy of 89% (std = 4.2%), an average TPR of 88.2% (std = 4.2%) and an average TNR of 89.6% (std = 4.9%) in distinguishing malignant and benign breast lesions (Table 3).

We further evaluated TNet and BNet models on an external data set of 102 unseen thyroid cases (57 benign and 45 malignant), and TNet model achieved an accuracy of 86.3%, with 84.4% and 87.7% for TPR and TNR respectively. Using the same set of thyroid US images, the BNet achieved a lower level of accuracy of 77.5% with 67.6% and 86% for TPR and TNR respectively. A BNet model trained on 321 benign images and 247 malignant images was tested on the external 104 breast cases (52 benign and 52 malignant), and the model achieved an accuracy of 87.5%, with 88.5% and 86.5% for TPR and TNR respectively.

Regarding the diagnostic performance, the TNet model achieved an AUC of 0.861 (95% CI: 0.792-0.929) in classifying malignant thyroid lesions which was comparable to that of the average performance of the three expert radiologists (0.810, 95%CI: 0.720-0.900) (Figure 4). The lowest AUC of the radiologists was 0.757 (95% CI:0.658-0.855), and the highest AUC was 0.854 (95% CI:0.775-0.934) (Table 4). The performance of three individual radiologists, however, was lower than that of the deep learning model in classifying thyroid cancer (radiologist 1 vs. TNet: p=0.0004; radiologist 2 vs. TNet: p=0.1536; radiologist 3 vs. TNet: p=0.0424). The results of each radiologist are provided in Table 5. Similar results were achieved in classifying malignant breast lesions in terms of sensitivity and accuracy rate. The TNet achieved higher sensitivity (88.5%) and accuracy rate (86.5%) than that of the three radiologists (sensitivity: 50.0% - 65.4%; accuracy: 71.2% - 78.8%) (Table 5). However, all of three radiologists had higher specificity (86.5% - 98.1%) than that of the TNet (84.6%). The results shown the effectiveness of our generic CNN model (TNet) to differentiate between malignant and benign breast lesions and thyroid lesions (Figure 5) compared with that of the radiologists.

4. Discussions

Our work provides additional support to the conclusions of previous studies that demonstrated deep learning algorithm performance comparable to radiologists or even better. For example, Han *et al.* developed a GoogLeNet-based model to distinguish between malignant and benign breast lesions with a large sample of 4254 benign lesions and 3,154 malignant lesions. The model achieved high sensitivity (86%), specificity (93%), and accuracy (91%) [22]. Guan *et al.* tested the ability of an inception-v3-based model to classify 1,275 papillary thyroid carcinomas and 1,162 benign lesions [23].

The model achieved sensitivity (93.3%), specificity (87.4%), and accuracy (90.5%). Ma *et al.* developed a pre-trained CNN model to predict of thyroid malignancy using 15,000 US images [24]. This model achieved a similar diagnostic performance as ours, with the sensitivity, specificity, and accuracy of their model as follows: 82.41% ± 1.35%, 84.96% ± 1.85%, and 83.02% ± 0.72%, respectively. Buda *et al.* produced a deep learning algorithm for thyroid cancer recognition based on 1,377 images that had a diagnostic performance similar to that of nine radiologists [9]. Specifically, their model achieved an AUC (0.87; 95% CI:0.76-0.95) that was comparable to that of nine skilled radiologists (0.82; 95% CI: 0.73-0.90) (p=0.38).

In a brief report on a separate study by Park *et al.* [11] with a large dataset, performances of two types of CAD systems (one using deep learning and the other support vector machine) were compared with those from experienced and inexperienced radiologists. The study found that the CAD systems had comparable performances to the radiologists. However, it was not clear from the report regarding which deep learning architecture was used or utilized, nor the selection of the radiologists taken part in the study. Wang *et al.* also conducted a large-scale study on multiple thyroid nodule classification [12]. Both Inception-ResNet-v2 and VGG-19 (chosen by this study) architectures were investigated. However, the image modality of the investigation was microscopic histological images rather than US images. Li *et al.* established a Faster R-CNN based method for distinguishing thyroid papillary carcinoma [25]. Their results demonstrated that the model improved the cancer classification over the manual methods but using a rather small dataset of 300 US images. In particular, the type of thyroid cancer was limited to thyroid papillary carcinoma in the study of Guan *et al.* and Li *et al.*, even though it is the most common primary thyroid cancer [25, 26]. The researchers, however, only designed one model for classifying either breast cancer or thyroid cancer. Liu et al proposed a multi-scale nodule detection scheme and a clinical-knowledge-guided CNN-based method to classify thyroid cancers [27]. By introducing clinical prior knowledge, such as margin, shape, aspect ratio, composition, and calcification, their results showed an impressive sensitivity of 98.2%, specificity of 95.1%, and accuracy rate of 97.1%. The method involves using three separate CNNs to extract features within the nodule boundary, around margin areas and between nodule and surrounding tissues. As a result, the architecture of the network is complex with a higher risk of model overfitting. Besides, all images were collected from US machines of a single make. None of the published work developed a consolidated algorithm to classify both breast and thyroid cancer.

In this paper, we developed a generic deep learning algorithm to classify thyroid and breast cancers with the following reasons. First, both cancers share common genetic features and are influenced by similar families of hormones [28,29]. For example, one study demonstrated the high frequency of thyroid stimulating hormone receptors in breast tissue [29]. Estrogen (which is highly expressed in breast tissue) might also contribute to thyroid gland development and pathology [30]. Furthermore, a common molecular mechanism may contribute to the concurrent thyroid cancer and breast cancers [31]. An *et al.* identified an increased risk of second primary carcinoma of the thyroid or breast in 6,833 patients with prior breast cancer or 4,243 patients with prior thyroid cancer [31]. Other factors such as increased thyroid peroxidase levels may also correlate with improved outcomes in patients with breast cancer [29]. In clinical practice, there was an elevated risk of developing a second primary cancer during the first year following the diagnosis of breast cancer [32]. These findings suggest that medical surveillance of breast cancer/thyroid cancer patients on the second primary cancer development is required.

To the best of our acknowledge, the work reported in this paper is the first to propose a generic CNN model (TNet) that showed a promising diagnostic performance in classifying both thyroid cancer and breast cancer. In the external test dataset, the TNet model distinguished benign and malignant breast lesions with a significantly higher sensitivity (88.5%) and accuracy rate (84.6%) without sacrificing too much on specificity (86.5%) than the radiologists (sensitivity: 50.0% - 65.4%; accuracy: 71.1% - 78.8%; and specificity: 86.5% - 98.0%). We used a higher percentage of malignant training data (44.5%) than the actual incidence rate (0.29%) [33], which might have rendered the algorithm more sensitive to malignant lesions, and therefore enabled a higher sensitivity than specificity. On the other hand, BNet showed a promising diagnostic performance in classifying thyroid cancer as well. It achieved a higher sensitivity (67.6%) and accuracy rate (77.5%) compared with that of the average performance of three radiologists (sensitivity: 57.7%, and accuracy: 75.0%), but a lower specificity (86%, the average performance of three radiologists: 92.3%). The BNet model also achieved comparable, and even marginally higher performances to the TNet on classifying the external breast cases. The results accord with previous studies, which showed that the application of machine learning in breast ultrasound achieved high level of differentiation between benign and malignant breast lesions, with an accuracy comparable to radiologists [34, 35].

Our work is primarily motivated by the interest in developing a generic CNN model suited for both thyroid and breast lesions given the similarity in the features of both types of lesions. Such approach could be useful when the data and annotation of one cancer type are not readily available. In order to explore the potentials of the generic approach for cancer diagnosis, we made a step further in building a CNN-based model on the same underlying DCNN architecture using combined cases of thyroid and breast lesions. We used 542 benign and 532 malignant RoI images of both types of lesions, and trained a new model TBNet with these images. We then tested the TBNet model on 204 cases (102 thyroid and 102 breast lesions). The overall accuracy was 82.3% with 74.4% sensitivity and 88.6% specificity. Again, the overall accuracy and sensitivity of TBNet seemed higher than those by the radiologists, and the specificity matched that by the radiologists. This initial trial test also shows the potentials of the generic approach for lesion classification.

A deep learning method to classify malignancy could contribute to clinical practice in different ways. First, multiple studies have confirmed that patients with previous breast or thyroid cancer have a significant increased overall risk of developing a secondary thyroid or breast cancer [36,37]. The TNet model could assist radiologists to screen both the thyroid gland and mammary gland of the same patient at the same time.

Consequently, the TNet model could improve the early detection rate. Second, deep learning methods produce consistent predictions for one given US image while predictions made by radiologists can vary depending on the individual level of experience and understanding. Finally, automated deep learning solutions can significantly reduce the image interpretation time in clinics. The readout time for the TNet model was around 1.15 seconds per image. By contrast, the radiologists took approximately 30-40 seconds to classify one thyroid/breast US image. For the external test dataset, three radiologists were asked to review images under time constraints in a real-life setting. The labor-intensive US image interpretation might well be one of the main reasons why the radiologists misclassified the malignant thyroid and breast lesions in the aforementioned results.

Some limitations of our study should also be noted. As a pilot study, our investigation confers the expected limitations of a retrospective and single center study with a limited number of samples. The proposed augmentation methods had to be used to enlarge the data sample sufficiently to train the CNN models. Furthermore, most patients involved in the study are southern Han Chinese. Nevertheless, the test results on the TNet model so far suggest that the model has the potential to perform better than skilled radiologists. We did ensure, however, that the US images included in the present study were obtained from different US machine makes. This helped ensuring data diversity for training more robust models.

5. Conclusion

In conclusion, the CNN-based models (TNet, BNet and even TBNet) have shown good performance in classifying both thyroid and breast cancers. The proposed generic deep learning framework can offer a promising diagnostic performance at classifying cancers of different types. For patients who are with thyroid or breast cancer history, such a consolidated model can lead to a more rapid intervention with the most appropriate treatment.

Encouraged by the results, we plan to expand the current research in several ways. Firstly, we will continue the ongoing investigation into the combined model TBNet by analyzing larger datasets collected from different centers involving diverse patient populations. Furthermore, a more systematic comparison between the models and radiologists of a wider range of experiences from several centers should be conducted under different control settings. We will also further analyze the relationship between a correct classification outcome made by the models and regions of input RoI images to identify the specific common features that the models have captured. Intrigued by the comparable performance of TNet and BNet on classifying breast lesions, we wish to investigate further the known ultrasound characteristics (e.g. shape ratio, hypo-echogenicity, and ill-defined margins) shared by thyroid and breast lesions. In addition, we will further investigate any new image textures learned by both models to identify potentially new common US characteristics useful for the diagnosis of thyroid and breast cancers.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians. American Cancer Society; 2018;68:394–424.

2. Hoang JK, Middleton WD, Farjat AE, Teefey SA, Abinanti N, Boschini FJ, et al. Interobserver variability of sonographic features used in the American College of Radiology Thyroid Imaging Reporting and Data System. AJR Am J Roentgenol; 2018;211:162-167.

3. Juri Yanase, Evangelos Triantaphyllou. A systematic survey of computer-aided diagnosis in medicine: past and present developments. Expert Systems with Applications. 2019 Dec 30;138, 112821

4. Tsantis S, Dimitropoulos N, Cavouras D, Nikiforidis G. Morphological and wavelet features towards sonographic thyroid nodules evaluation. Comput Med Imaging Graph. 2009 Mar;33: 91-9.

5. Keramidas EG, Lakovidis DK, Maroulis D, Dimitropoulos N. THyroid texture representation via noise resistant image features. In: Proceedings of the IEEE Symposium on Computer-Based Medical Systems, 560-565. 2008

6. Song G, Xue F, Zhang C. A model using texture features to differentiate the nature of thyroid nodules on sonography. J Ultrasound Med 2015 Oct;34:1753-60

7. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60-88

8. Deng J, Dong W, Socher R, Li L-J, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition [Internet]. Miami, FL: IEEE; 2009 [cited 2019 Jun 7]. page 248–55. Available from: https://ieeexplore.ieee.org/document/5206848/

9. Buda M, Wildman-Tobriner B, Hoang JK, Thayer D, Tessler FN, Middleton WD, et al. Management of thyroid nodules seen on US images: deep learning may match performance of radiologists. Radiology. 2019 Jul 9:181343 [Epub ahead of print].

10. Barinov, L, Jairaj, A, Paster, L, Hulbert, W, Podilchuk, C. Decision Quality Support in Diagnostic Breast Ultrasound through Artificial Intelligence. The Science Education and Research Center at Temple University The IEEE Signal Processing in Medicine and Biology Symposium (SPMB16). IEEE. 2016

11. Park VY, Han K, Seong YK, Park MH, Kim EK, Moon HJ, et al. Diagnosis of thyroid nodules: performance of a deep learning convolutional neural network model vs. radiologists. Sci Rep. 2019 Nov 28;9(1):17843.

12. Wang Y, Guan Q, Lao I, Wang L, Wu Y, Li D, et al. Using deep convolutional neural networks for multi-classification of thyroid tumor by histopathology: a large-scale pilot study. Ann Transl Med. 2019 Sep;7(180):468.

13. Sandeep TC, Strachan MW, Reynolds RM, Brewster DH, Scelo G, Pukkala E, et al. Second primary cancers in thyroid cancer patients: a multinational record linkage study. J Clin Endocrinol Metab. 2006;91:1819-25.

14. Tanaka H, Tsukuma H, Koyama H, Kinoshita Y, Kinoshita N, Oshima A. Second primary cancers following breast cancer in the Japanese female population. Jpn J Cancer Res. 2001;92:1-8.

15. Nielsen SM, White MG, Hong S, Aschebrook-Kilfoy B, Kaplan EL, Angelos P, et al. The breast-thyroid cancer link: a systematic review and meta-analysis. Cancer Epidemiol Biomarkers Prev. 2016;25:231-8.

16. Melany M. Ultrasound Imaging of Thyroid Cancer. In: Braunstein GD, editor. Thyroid Cancer [Internet]. Boston, MA: Springer US; 2012 [cited 2019 Jun 7]. page 63–91. Available from: http://link.springer.com/10.1007/978-1-4614-0875-8_4

17. Sencha AN, Evseeva EV, Mogutov MS, Patrunov YN. Ultrasound Diagnosis of Breast Cancer. Breast Ultrasound [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013 [cited 2019 Jun 7]. page 49–122. Available from: http://link.springer.com/10.1007/978-3-642-36502-7_4

18. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:14091556 [cs] [Internet]. 2014 [cited 2019 Jun 7]; Available from: http://arxiv.org/abs/1409.1556

19. Tessler FN, Middleton WD, Grant EG. Thyroid Imaging Reporting and Data System (TI-RADS): A User's Guide. Radiology. 2018;287(3):1082.

20. Mercado CL. BI-RADS update. Radiol Clin North Am. 2014;52:481-7.

21. Shorten, Connor, and Taghi M. Khoshgoftaar. "A survey on image data augmentation for deep learning." Journal of Big Data 6, no. 1 (2019): 60.

22. Han S, Kang HK, Jeong JY, Park MH, Kim W, Bang WC, et al. A deep learning framework for supporting the classification of breast lesions in ultrasound images. Phys Med Biol. 2017;62:7714-28.

23. Guan Q, Wang Y, Du J, Qin Y, Lu H, Xiang J, et al. Deep learning based classification of ultrasound images for thyroid nodules: a large scale of pilot study. Ann Transl Med. 2019;7;137.

24. Ma J, Wu F, Zhu J, Xu D, Kong D. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. Ultrasonics. 2017;73:221-30.

25. Li H, Weng J, Shi Y, Gu W, Mao Y, Wang Y, Liu W, et al. An improved deep learning approach for detection of thyroid papillary cancer in ultrasound images. Sci Rep. 2018 Apr 26;8(1):6600.

26. Miccoli P, Bakkar S. Surgical management of papillary thyroid carcinoma: an overview. Updates Surg. 2017;69:145-50.

27. Liu T, Guo Q, Lian C, Ren X, Liang S, Yu J, et al. Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks. Med Image Anal. 2019 Dec;58:101555.

28. Agarwal DP, Soni TP, Sharma OP, Sharma S. Synchronous malignancies of breast and thyroid gland: a case report and review of literature. J Cancer Res Ther. 2007;3:172-3.

29. Turken O, Narin Y, Demlrbas S, Onde ME, Sayan O, Kandemlr EG, et al. Breast cancer in association with thyroid disorders. Breast Cancer Res. 2003;5:R110-3.

30. Kawabata W, Suzuki T, Moriya T, Fujimori K, Naganuma H, Inoue S, et al. Estrogen receptors (alpha and beta) and 17beta-hydroxysteroid dehydrogenase type 1 and 2 in thyroid disorders: possible in situ estrogen synthesis and actions. Mod Pathol. 2003;16:437-44.

31. An JH, Hwangbo Y, Ahn HY, Keam B, Lee KE, Han W, et al. A possible association between thyroid cancer and breast cancer. Thyroid. 2015;25:1330-8.

32. Tanaka H, Tsukuma H, Koyama H, Kinoshita Y, Kinoshita N, Oshima A. Second primary cancers following breast cancer in the Japanese female population. Jpn J Cancer Res. 2001;92:1-8.

33. Li T, Mello-Thoms C, Brennan PC. Descriptive epidemiology of breast cancer in China: incidence, mortality, survival and prevalence. Breast Cancer Res Treat. 2016;159:395-406.

34. Becker AS, Mueller M, Stoffel E, Marcon M, Ghafoor S, Boss A. Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. Br J Radiol. 208;91(1083):20170567.

35. Fleury E, Marcomini K. Performance of machine learning software to classify breast letions using BI-RADS radiomic features on ultrasound images. Eur Radiol Exp. 209:3:34.

36. Dobrinja, C, Scomersi, S, Giudici, F, Vallon, G, Lanzaro, A, Troian, M, et al. Association between benign thyroid disease and breast cancer: a single center experience. BMC Endocr Disord. 2019;19:104.

37. Dong L, Lu J, Zhao B, Wang W, Zhao Y. Review of the possible association between thyroid and breast carcinoma. World J Surg Oncol. 2018;16:130.

Tables

Table 1: Study population with breast lesions and baseline characteristics

	Training		Testing	
	Malignant	Benign	Malignant	Benign
Patients (years old)*	60.3 ± 11.7	55.3 ± 12.6	65.7 ± 15.1	59.3 ± 10.8
Number of lesions	299	373	52	52
Planes of US images				
Longitudinal	176	251	27	28
Transverse	123	122	25	24
US machine types				
Philips	138	206	19	32
GE	76	83	10	8
Toshiba	43	50	5	6
Siemens	42	34	18	6
BI-RADS				
2	0	149	0	27
3	4	125	0	8
4a	127	75	30	11
4b	65	23	5	6
4c	42	1	7	0
5	61	0	10	0

Table 2 Study population with thyroid lesions and baseline characteristics

	Training		Test	
	Malignant	Benign	Malignant	Benign
Patients (years old)*	58.5 ± 10.4	54.2 ± 8.1	55.8 ± 10.9	53.9 ± 7.3
Number of lesions	298	421	45	57
Location				
Right	150	198	29	27
Left	138	196	8	18
Isthmus	10	27	8	12
Planes of US images				
Longitudinal	211	314	31	40
Transverse	87	107	14	17
US machine types				
Philips	155	198	23	27
GE	58	107	8	11
Toshiba	37	55	9	5
Siemens	48	61	5	14
TI-RADS				
2	0	187	0	32
3	11	136	0	11
4a	126	68	31	9
4b	89	30	6	4
4c	35	0	3	1
5	37	0	5	0

*The data represent the means ± standard deviation.

Table 3 Average TPR, TNR, accuracy and AUC for 10 folds for both TNet and BNet

Models	Evaluation Measurements			
	TPR (std)	TNR (std)	Accuracy (std)	Mean AUC
TNet	83.9% (3.9%)	88.6% (4.6%)	86.5% (2.8%)	0.863
BNet	88.2% (4.2%)	89.6% (4.9%)	89% (4.2%)	0.888

Table 4: Diagnostic performance of the TNet model and radiologists

Thyroid	AUC	95% CI	Breast	AUC	95% CI
TNet	0.861	0.792-0.929	TNet	0.875	0.804-0.947
AvgR	0.810	0.720-0.900	AvgR	0.750	0.653-0.847
R1	0.757	0.658-0.855	R1	0.756	0.660-0.853
R2	0.854	0.775-0.934	R2	0.698	0.593-0.802
R3	0.830	0.744-0.916	R3	0.777	0.682-0.872

R1-R3 indicates radiologists 1 to 3. AvgR indicates the average performance of the three radiologists.

Table 5 TPR, TNR, and accuracy of TNet and the three radiologists

	TNet			Radiologist 1			Radiologist 2			Radiologist 3		
	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR	ACC
Thyroid	84.4%	87.7%	86.3%	68.9%	82.5%	76.5%	86.7%	84.2%	85.3%	80.0%	86.0%	83.3%
Breast	88.5%	84.6%	86.5%	65.4%	86.5%	76.0%	50.0%	92.3%	71.2%	59.6%	98.1%	78.8%

TPR indicates true positive rate. TNR indicates true negative rate. ACC indicates accuracy rate.

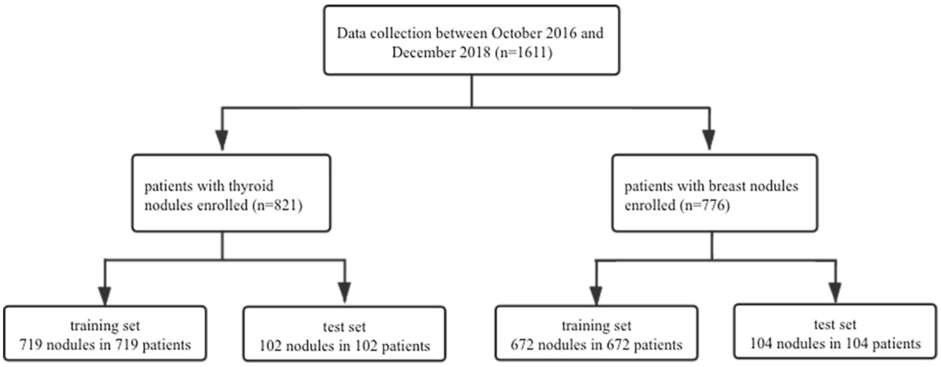
Data collection between October 2016 and December 2018 (n=1611)
patients with thyroid nodules enrolled (n=821)
patients with breast nodules enrolled (n=776)
training set 719 nodules in 719 patients
test set 102 nodules in 102 patients
training set 672 nodules in 672 patients
test set 104 nodules in 104 patients

Figure 1: Flowchart of the study population in the training and testing sets.

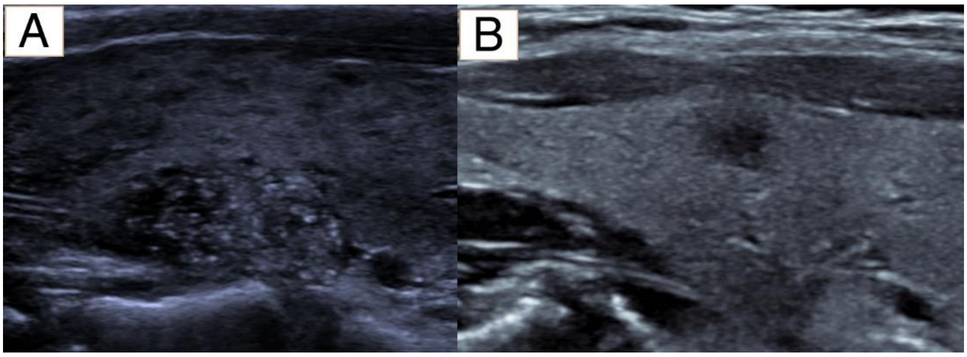

Figure 2: Representative US images showing malignant thyroid lesions.

(a) A malignant wider-than-tall, solid lesion with punctate echogenic foci. All radiologists and the TNet model correctly classified the lesion.

(b) A malignant wider-than-tall, hypoechoic solid lesion with an ill-defined margin. All radiologists misclassified the lesion as benign due to the small size of the lesion (0.8cm) and no punctate echogenic foci while the TNet model correctly classified the lesion as malignant.

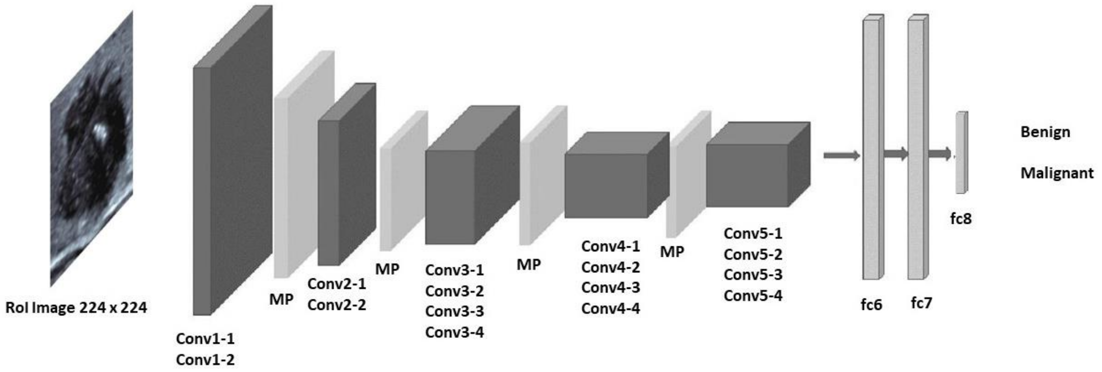

Figure 3: CNN architecture consists of 16 convolutional (Conv) layers with 3x3 kernels with depths 64, 128, 256, 512 for Conv1, Conv2, Conv3, Conv4 and Conv5, respectively; max pooling layers (MP) and 3 fully connected (fc) layers fc6, fc7 and fc8 with sizes 4096, 4096 and 2, respectively.

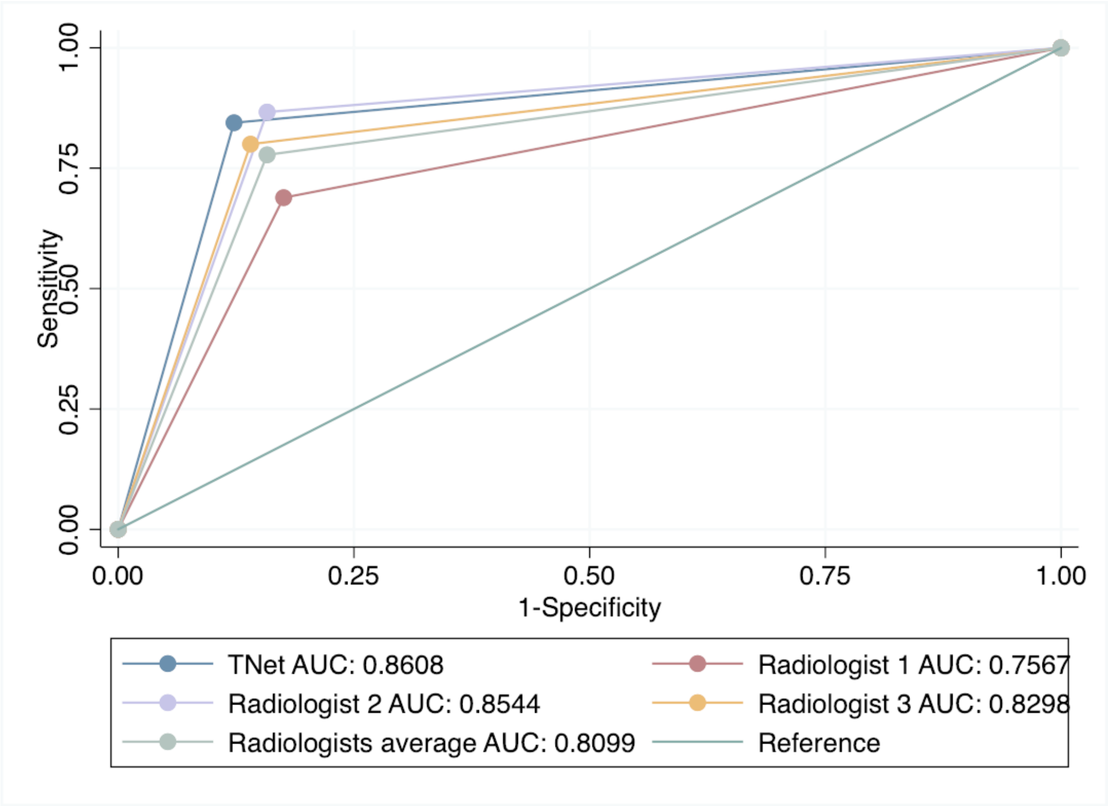

Figure 4: ROC curves for binary classification revealing diagnostic performances of TNet, 10-fold cross validation TNet, and three radiologists.

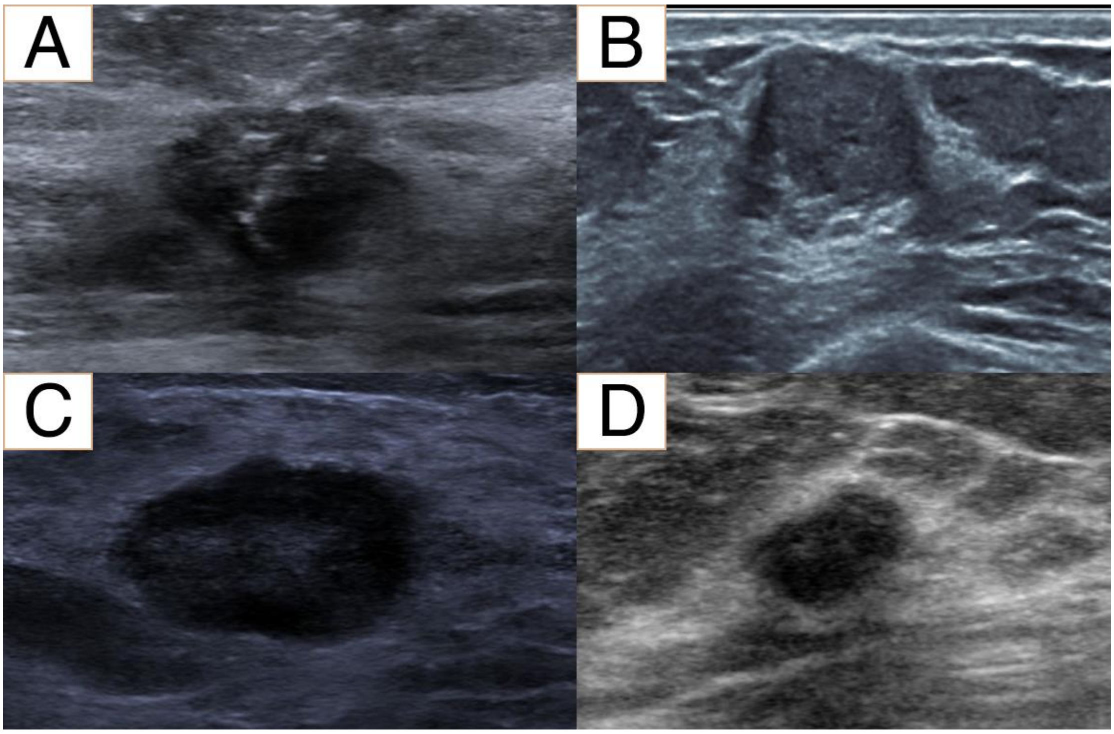

Figure 5: Representative US images showing malignant breast lesions.

(a) A malignant lesion with irregular shape, calcification, and not circumscribed margin. All radiologists and the TNet model correctly classified the lesion.

(b) A malignant lesion with an oval shape, circumscribed margins, and enhancement posterior features. All radiologists and the TNet model misclassified the lesion as benign due to the enhancement posterior features that result in a soft tissue.

(c) A hypoechoic malignant lesion. All radiologists correctly classified the lesion as malignant, while the TNet model misclassified the lesion as benign.

(d) A heterogeneous, hypoechoic lesion with an oval shape and parallel orientation characteristic of malignant lesions. All radiologists misclassified the lesion as benign due to the small size of the lesion (1.4cm) and parallel orientation, while the TNet model correctly classified the lesion as malignant.