Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

A new xAI framework with feature explainability for tumors decision-making in Ultrasound data: comparing with Grad-CAM

Di Song[a,†], Jincao Yao[b,†], Yitao Jiang[c,†], Siyuan Shi[d], Chen Cui[d], Liping Wang[b], Lijing Wang[b], Huaiyu Wu[a], Hongtian Tian[a], Xiuqin Ye[a], Di Ou[b], Wei Li[b], Na Feng[b], Weiyun Pan[b], Mei Song[b], Jinfeng Xu[a,*], Dong Xu[b,*], Linghu Wu[a,*], Fajin Dong[a,*]

[a] *Department of Ultrasound, Shenzhen People's Hospital, The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology, Shenzhen 518020, Guangdong, China*
[b] *The Cancer Hospital of the University of Chinese Academy of Sciences (Zhejiang Cancer Hospital), Institute of Basic Medicine and Cancer (IBMC), Chinese Academy of Sciences, Hangzhou, Zhejiang 310022, China*
[c] *Research and development department, Microport Prophecy, Shanghai 201203, China*
[d] *Research and development department, Illuminate, LLC, Shenzhen, Guangdong 518000, China*

ARTICLE INFO

Article history:
Received 15 August 2022
Revised 13 March 2023
Accepted 2 April 2023

Keywords:
Explainability
Artificial intelligence
Medical decision support
Ultrasound
Thyroid cancer

ABSTRACT

Background and objective: The value of implementing artificial intelligence (AI) on ultrasound screening for thyroid cancer has been acknowledged, with numerous early studies confirming AI might help physicians acquire more accurate diagnoses. However, the black box nature of AI's decision-making process makes it difficult for users to grasp the foundation of AI's predictions. Furthermore, explainability is not only related to AI performance, but also responsibility and risk in medical diagnosis. In this paper, we offer Explainer, an intrinsically explainable framework that can categorize images and create heatmaps highlighting the regions on which its prediction is based.

Methods: A dataset of 19341 thyroid ultrasound images with pathological results and physician-annotated TI-RADS features is used to train and test the robustness of the proposed framework. Then we conducted a benign-malignant classification study to determine whether physicians perform better with the assistance of an explainer than they do alone or with Gradient-weighted Class Activation Mapping (Grad-CAM).

Results: Reader studies show that the Explainer can achieve a more accurate diagnosis while explaining heatmaps, and that physicians' performances are improved when assisted by the Explainer. Case study results confirm that the Explainer is capable of locating more reasonable and feature-related regions than the Grad-CAM.

Conclusions: The Explainer offers physicians a tool to understand the basis of AI predictions and evaluate their reliability, which has the potential to unbox the "black box" of medical imaging AI.

1. Introduction

Thyroid cancer is one of the most common endocrine cancers, with a rapidly rising incidence, and has risen to become the fifth most common cancer in adult women worldwide [1,2]. Thyroid nodules are an indicator of thyroid cancer, which is an abnormal region of the thyroid in adults [3]. Although most thyroid nodules are benign, it is essential to detect them at an early stage and monitor their growth to evaluate the necessity of clinical intervention [4]. Ultrasound is one of the most frequently used modalities for screening thyroid diseases, with highly operator-dependent accuracy [5]. However, since ultrasound screening is conducted by physicians with varying degrees of experience, the issue of nodule malignancy diagnosis is far from objective. According to previous research, one main problem is the low specificity of human diagnosis [6].

Artificial intelligence (AI) has been widely used to assist clinical diagnosis in the medical field, including thyroid disease di-

[*] Corresponding authors.
E-mail addresses: songdi@szhospital.com (D. Song), yaojc@zjcc.org.cn (J. Yao), joetao097@gmail.com (Y. Jiang), siyuanshinu@gmail.com (S. Shi), antony.cuichen@gmail.com (C. Cui), wanglp@zjcc.org.cn (L. Wang), wanglj844@zjcc.org.cn (L. Wang), 234721800@qq.com (H. Wu), tianhongtian@szhospital.com (H. Tian), oudi@zjcc.org.cn (D. Ou), liwei@zjcc.org.cn (W. Li), fengna@zjcc.org.cn (N. Feng), panwy@zjcc.org.cn (W. Pan), songmei@zjcc.org.cn (M. Song), xujinfeng@yahoo.com (J. Xu), xudong@zjcc.org.cn (D. Xu), wulinghu@szhospital.com (L. Wu), dongfajin@szhospital.com (F. Dong).

[†] These authors contributed equally to this work.

https://doi.org/10.1016/j.cmpb.2023.107527
0169-2607/© 2023 Elsevier B.V. All rights reserved.

agnosis. Buda et al. constructed a convolutional neural network for benign-malignant classification of thyroid nodules, achieving equivalent performance as human experts with 87% sensitivity and 52% specificity [7]. Liu et al. proposed a system that can not only predict malignancy but also auto-detect various nodular features based on the thyroid imaging reporting and data system (TI-RADS) [8]. A previous reader study suggests AI has similar sensitivity and improved specificity compared with the judgment of experienced radiologists [9]. To summarize, AI could be a practical method to address the issue of low specificity in thyroid ultrasound diagnosis.

The potential of AI in thyroid diagnosis has already been recognized, however, it suffers from a lack of transparency in the decision-making process. Since the complex nature of deep learning, it is impossible to know exactly what features are learned by neural networks, thus making it difficult to explain what the AI predictions' basis is [10,11]. Meanwhile, the medical-decision-support system is related to human life and requires transparency. The issue of interpretability is not only a matter of theoretical development, but also a core consideration for risk and responsibilities [11–13]. One common scenario is when physicians' initial diagnoses are opposed by AI, and the physician must decide whether to believe the AI's judgment. Without a trustworthy explanation module, physicians risk being misled and making incorrect diagnoses. In the medical domain, for example, AI can make unreliable decisions in safety-critical scenarios [14]. It is essential to unpack the black-box of AI and justify its reliability, robustness, and explainability [15]. As explainability and robustness can promote reliability and trust and ensure that humans remain in control, AI can supplement human intelligence [16].

Explainable AI (XAI) has been an emerging topic of machine learning research, especially in medicine. Arrieta et al. categorized XAI into model-specific and model-agnostic, both of which can be either intrinsic or post-hoc [17]. One philosophy of static image XAI is that a high-performance deep learning model actually identifies the appropriate area of the image and does not overemphasize unimportant findings [11]. Several methods for creating visual explanations for prediction from image classification models have been developed so far, including occlusion maps, salience maps, class activation maps, and attention maps [18–21]. The majority of current visual explanation methods are based on post hoc techniques. Post hoc techniques start with the deep model's output and go backwards into the model using different computer algorithmic tools to generate meaningful explanations [22], such as Gradient-weighted Class Activation Mapping (Grad-CAM), which is a model-agnostic variation of Class Activation Maps (CAM) that does not require a specific global average pooling layer [18]. Although post-hoc methods provide information on which regions are important to the model, they do not provide a precise answer to the question of what the basis of the AI predictions is [17,23,24], whereas intrinsic methods do. For example, occlusion sensitivity systematically occludes different portions of the input image with a gray square to identify the true regions where the AI's prediction is based [20]. The main flaws of occlusion sensitivity are that the size of the square is a hyperparameter that must be manually set, and the systematically occlude process takes time.

XAI assistance in achieving trustworthy and robust medical AI-based systems and the use of new types of human-AI interfaces and supportive visualizations [25]. The purpose of this study was to propose a novel framework, called the Explainer, for thyroid benign-malignant classification and TI-RADS feature classification (margin, composition, echogenicity, and echogenic foci), as well as to test the performance and reliability of the Explainer.

2. Materials and Methods

In this section, we describe the procedure for using the Explainer on image classification CNNs. The Explainer is a tiny neural network that is parasitic on the CNNs and is used to explain what the AI predictions' basis is.

The architecture of the Explainer is depicted in Fig. 1, which contains a pretrained CNN model (a CNN backbone cascaded by a CNN classifier), and an Explainer generator. The Explainer generator explains the pretrained CNN model. The CNN backbone extracts significant feature maps from the image input, and the fea-

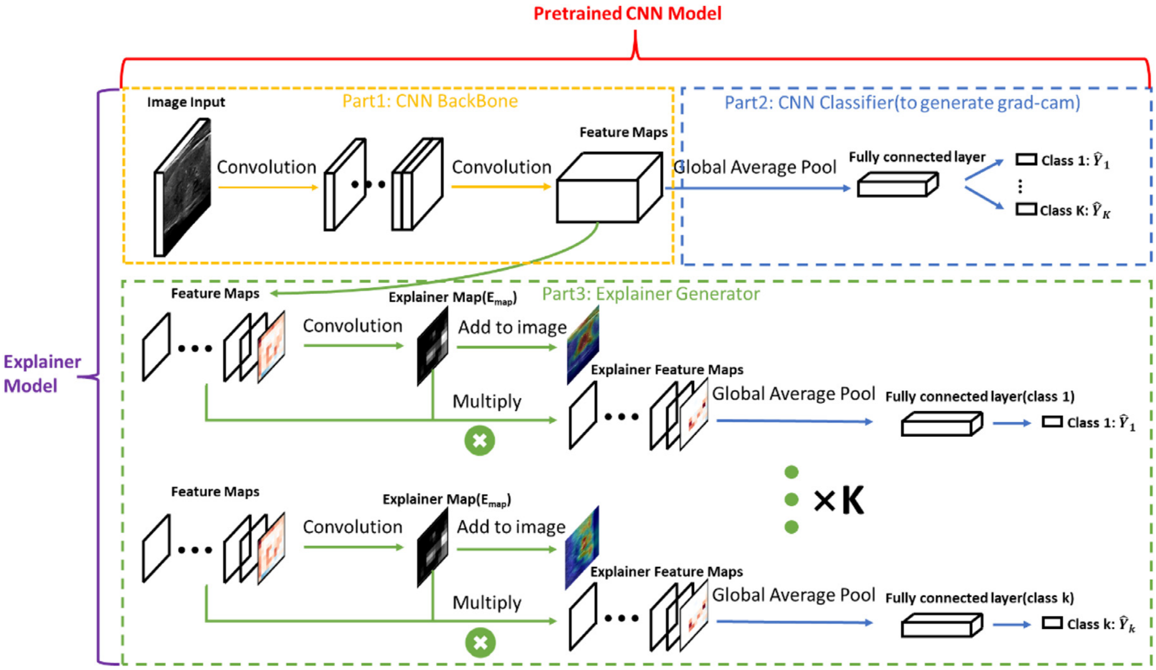


Fig. 1. Architecture of the Explainer.

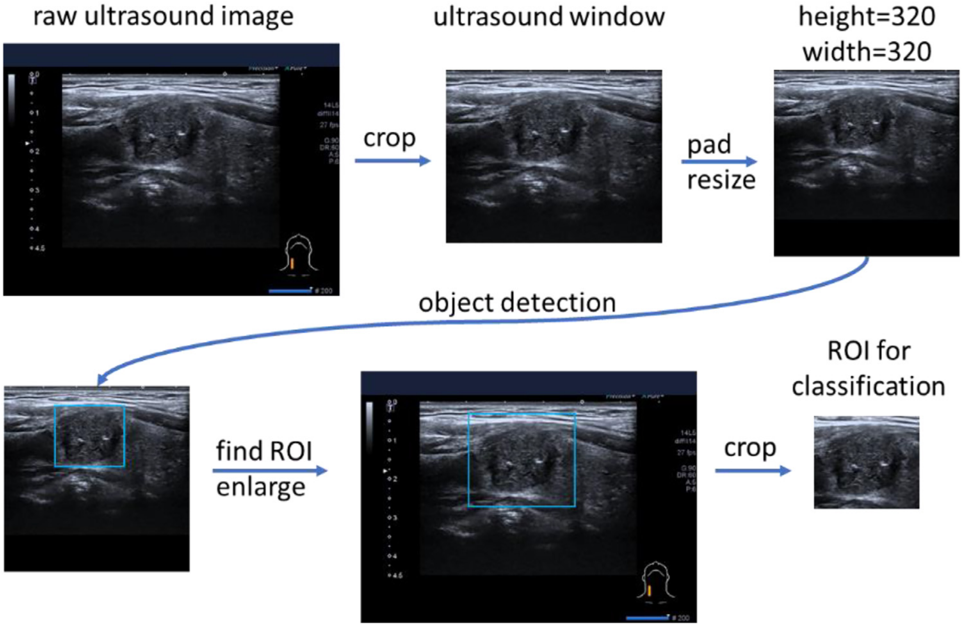


Fig. 2. ROI extraction pipeline.

ture maps are subsequently sent to the CNN classifier and the Explainer generator. The CNN classifier can be used to generate the Grad-CAM for the input image, which serves as the true mask for the Explainer generator training. The Explainer generator has two functions: the first is to generate an E_{map} from the convolution of feature maps, which highlights the region that the CNN model uses to make its decision, and the second is to multiply the E_{map} on feature maps and utilize the resulting feature map for classification.

A CNN backbone, a CNN classifier, and an Explainer generator constitute an entire architecture. Only the green arrow convolution layers in the Explainer generator needs to be trained, while the rest of the layers in the Explainer inherit the weights from the pretrained CNN model.

2.1. Thyroid lesion detection and image cropping

One strategy we employed in our research is to perform an object detection task on the ultrasound images and pinpoint the lesion as shown in Fig. 2. The purpose of this method is to concentrate the model's attention to the lesion and tissue enclosing it, which simplifies the subsequent work. To train the object detection model (RetinaNet, [26]) we utilized 300 ultrasound pictures (150 horizontal and 150 vertical views). The input to the RetinaNet are the preprocessed ultrasound windows cropped from raw ultrasound images, and the output are the bounding boxes of the lesion. The preprocessing procedure includes padding the cropped window to square and resizing to a 320*320 shape. The mean average precision of the model is 0.9 given an Intersection over Union (IOU)threshold of 0.5. We increase the height and width of the detected box by 1.5 times to cover more tissue surrounding the nodule and crop the ROI from the raw ultrasound image. About 95% of the cropped regions can show a full extent of lesion. All images in data set are preprocessed, measured by the object detection model, and cropped for future usage.

We utilize object detection model to extract ROI of thyroid ultrasound images to concentrate the model's attention to the lesion and tissue enclosing it.

2.2. CNN model pretraining

A CNN model is consisted of a backbone for extracting image features and a classifier for making prediction. Most CNNs can serve as the backbone model for the Explainer, such as the DenseNet [27], the ResNet [28], the Xception [29], the MobileNet [30] and so on. The dataset is separated into three subsets for training, validation, and testing, a 6:1:3 splitting rate is utilized. Image augmentation, such as rotation, zoom, translation, flip, and grayscale modification, are used to help model generalization. The DenseNet-121 is used as the CNN (Fig. 1 part 1 and 2) and is trained from random initialization using the Adam optimizer to classify benign and malignant nodules. We call this model the pretrained CNN model.

2.3. Train the Explainer by using heatmaps generated by Grad-CAM

The purpose of the Explainer is to highlight the regions where the model relies on to make predictions. We utilize an E_{map}, which is a heatmap created by the Explainer generator, to highlight the regions. The E_{map} has the same width and height as the Explainer generator's input feature map. The values of the E_{map} are between 0 to 1, representing the importance for each pixel in feature maps.

We trained the parameters of the Explainer generator by supervised learning using both classification and segmentation tasks. Classification labels were determined according to pathological results or agreement among physicians. Segmentation labels were obtained by resizing the Grad-CAM to the shape of the Explainer generator's output, since the Grad-CAM could indicate the regions that might be essential for prediction. We used the Grad-CAM for generating segmentation labels since it has a wide range of applications on different CNN models [18]. The Grad-CAM is designed to use the target concept's gradients, which comes from the CNN classifier, to generate a heatmap that highlights significant places in the original image for prediction.

Although the Explainer uses the Grad-CAM as the ground truth (GT), the E_{map} explains the model in a completely different way. The E_{map} worked directly on feature maps by multiplying, where the pixel that the E_{map} deems unimportant is multiplied by 0 and

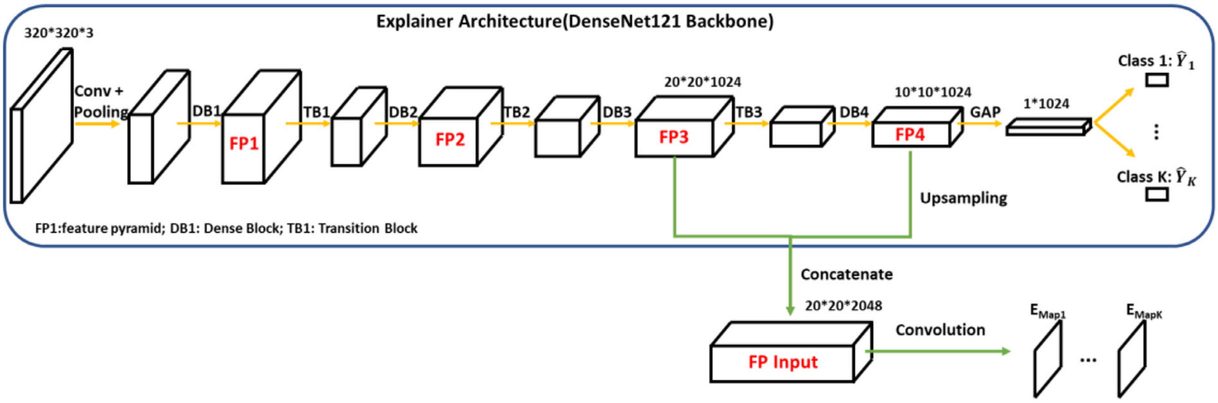


Fig. 3. Apply the Explainer generator on feature maps from the feature pyramid 3 and 4 of the DenseNet-121.

the feature represented by this pixel is not provided to the classifier. The Explainer and the Grad-CAM vary in that the Grad-CAM addresses the issue of which features the model considers significant, whilst the Explainer focuses on which features the model uses to make predictions. Features that a model considers important and features that the model uses to make predictions may overlap but are not exactly the same.

By reshaping the feature maps from separate layers into the same size, the Explainer generator can combine inputs from multiple layers, not limited to the last convolution layer. We applied the Explainer generator on feature maps from the last and the second last dense blocks of the DesneNet-121. The procedure is depicted in Fig. 3.

In feature pyramid 4 (FP4), feature maps have a 10*10 shape, but in feature pyramid 3 (FP3), they have a 20*20 shape. To reformat the feature maps into the same size, we used the upsampling operation on feature maps from the FP4. DB: dense block, TB: transition block, FP: feature pyramid, GAP: global average pooling.

2.4. Multitask loss is used in the Explainer

The total loss (L_{total}) of the Explainer is divided into two parts, a classification loss between the label (y) and the prediction (p), and a segmentation loss between the Grad-CAM ($\mathrm{G_{map}}$) and the $\mathrm{E_{map}}$. We used the Focal Loss for the classification task ($L_{fl}(y,p)$) and the Binary Cross-entropy Loss for the segmentation task ($L_{BCE}(G_{map_i},\ E_{map_i})$).

$$\boldsymbol{L_{fl}}\ (\boldsymbol{y}, \boldsymbol{p}) = -\boldsymbol{\alpha_t}(1-\boldsymbol{p_y})^{\boldsymbol{\gamma}}\log(\boldsymbol{p_y}),\ \ \boldsymbol{\alpha_t} = 0.5,\ \ \boldsymbol{\gamma} = 2 \tag{1}$$

where $y \in \{1,\ \ldots,\ k\}$ is an integer class label (suppose the classifier outputs k classes), and $p = (p_1, \ldots,\ p_k) \in [0,\ \ 1]^k$ is a vector that represents the model's predicted probability distribution over k classes.

$$L_{BCE}(G_{map_i},\ E_{map_i}) = \ -[G_{map_i} log E_{map_i} + (1 - G_{map_i})\log(1 - E_{map_i})] \tag{2}$$

where $G_{map_i} \in \{0, 1\}^{(height,\ width)}$ is a 2-dimensional matrix of zeros and ones that represents the $\mathrm{G_{map}}$ of the $\mathrm{i^{th}}$ class, and $E_{map_i} \in [0,\ \ 1]^{(height,width)}$ is a 2-dimensional matrix that represents the model's predicted probability for each of the output pixel of the $\mathrm{i^{th}}$ class.

As shown in Fig. 4, suppose the classifier outputs k classes, the Explainer generates $\mathrm{E_{map}}$ for all k classes, but the segmentation loss is calculated only for the pretrained CNN model's predicted class. The total loss is defined as

$$L_{total}\left(y, y', p, G_{map}, E_{map}\right) = \alpha L_{fl}(\boldsymbol{y}, \boldsymbol{p}) + \beta L_{BCE}\left(E_{map_{y'}}, G_{map_{y'}}\right) \tag{3}$$

where $y' \in \{1,\ \ldots,\ k\}$ is the integer class label predicted by the pretrained CNN model.

In Eq. (3), α and β represents the weight of the classification loss and the segmentation loss, respectively. We set $\alpha = 1$, $\beta = 0.1$ in our experiments.

p represents the model's predicted probability distribution over k classes. y' is a one-hot prediction over k classes according to the pretrained CNN model. E_{map} is the Explainer map generated for all classes. G_{map_i} is the Grad-CAM for class i if $y'_i = 1$, G_{mapi} is set to zero for class i if $y'_i = 0$.

2.5. Inherit the pretrained CNN model weight and train the Explainer convolution layers

The Explainer inherits both the structure and the weight of a pretrained CNN model. The trainable parameter in the Explainer is the convolution layer in the Explainer generator, which is shown by the green convolution arrow in Fig. 1.

We freeze all layers except the Explainer's convolution layer during training, which might be regarded a tiny parasite neural network. To monitor the training process, we utilized the loss and accuracy metrics for both the classification and segmentation. Since layers in the backbone which serve as a feature extractor have been trained, the loss may stop decreasing in a few epochs; in our experiment, training the Explainer with the DenseNet-121 backbone took 10 epochs.

2.6. The Explainer acts as a comprehensive predictor and explains itself

After the Explainer has been trained, the CNN classifier (part 2 in Fig. 1) could be deleted. The CNN backbone together with the Explainer generator constitutes a CNN model that can classify its input and output the $\mathrm{E_{map}}$. In our experiment, the Explainer has almost the same classification performance as the pretrained CNN model. Overlaying the $\mathrm{E_{map}}$ of the predicted class on the corresponding input image aids in explaining the model intuitively. The heatmap is reshaped to match the shape of the input image, then the OpenCV library [31] is used to merge the two.

3. Experiment Design

3.1. Data sources and entry criteria

This retrospective study was carried out in compliance with the protocols established by the participating hospitals. This research was approved by the Ethics Committee of the Cancer Hospital of The University of Chinese Academy of Sciences and Shenzhen People's Hospital, and each patient signed an informed consent form.

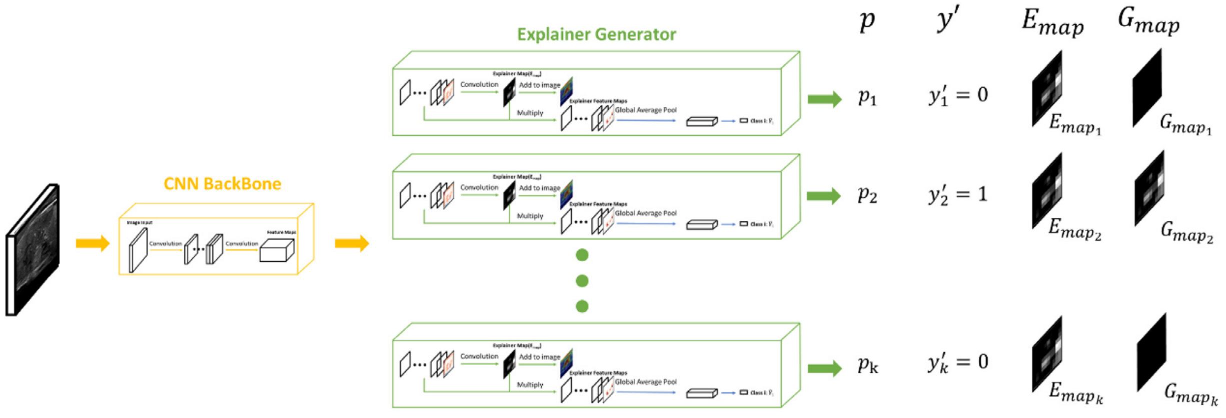


Fig. 4. Calculating the loss of the Explainer.

Table 1
The study population and images in training set, validation set, and test set. Bn: benign. Mal: malignant.

Set		Train		Validation		Test		Total		
Subject		4388		686		2162		**7236**		
Malignancy		Bn	Mal	Bn	Mal	Bn	Mal	Bn	Mal	All
Nodule		2709	1972	404	323	1372	934	4485	3229	**7714**
Images	Horizontal view	3901	1892	562	315	1865	637	6327	2844	**9171**
	Vertical view	4067	2313	587	375	1961	866	6616	3554	**10170**

The following criteria apply to data inclusion:

(1) The patient needs to undergo total thyroidectomy or subtotal thyroidectomy;
(2) Thyroid ultrasound examination is required before surgery, and there are available ultrasound images. The ultrasound image data of thyroid nodules should be clear and patient information should be complete;
(3) The first operation for thyroid disease. No history of previous neck surgery, including thyroid and cervical lymph node surgery;
(4) The clinical data of the patient are complete, including age, gender, past thyroid-related medical history, etc.

The following criteria apply to data exclusion:

(1) Preoperative thyroid ultrasound examination is missing or lacks a clear ultrasound image or the thyroid tumor is too large to obtain a complete image of the tumor;
(2) The pathological diagnosis is not clear;
(3) Patients with two or more neck surgeries;
(4) The clinical data of the patients were incomplete.

3.2. Study population

The study included 19341 2D ultrasound images (9171 horizontal views and 10170 vertical views) of 7714 nodules obtained from 7236 individuals from October 2019 to May 2021, as indicated in Table 1.

Each recruited individual has at least one nodule, and each nodule contains a maximum of four image records, including horizontal view raw data, horizontal view with markers, vertical view raw data, and vertical view with markers, some may not have all 4 records. We split the dataset into three sets based on patient ID: training, validation, and test, randomly with a 6:1:3 ratio. This would ensure that images from the same patient are in the same set. The collection includes 12943 images of benign nodules (6327 horizontal views and 6616 vertical views) and 6398 images of malignant nodules (2844 horizontal views and 3554 vertical views). Details are provided in Table 1.

All images were reviewed by at least three radiologists who were blind to the pathologic findings. Readers were given all available views of each nodule, including horizontal view and vertical views. The readers scored features for nodule composition, echogenicity, margins, and echogenic foci using the ACR TI-RADS standard. For composition, echogenicity, and echogenic foci, we collected each nodule's readings and assigned scores based on the majority result. If no majority result was obtained, the score was left blank. Because inter-observer agreement for scoring margin is low, we opted to designate a nodule's margin as abnormal if any reader's score is larger than zero, and normal if all readers scored it as zero. The shape score was calculated by comparing the diameters measured during the diagnostic process. Annotation results were shown in Table 2, N/A means the score is not available according to ACR TI-RADS.

As shown in Table 3, we generated the dataset for five independent deep learning classification tasks, including classifying malignancy, composition, echogenicity, margin, and echogenic foci, using the train-test splitting described previously.

3.3. Reader studies experiment design

We conducted a benign-malignant classification study to whether physicians perform better under the assistance of Explainer than diagnose alone or with Grad-CAM.

Complete physician diagnosis (Complete-Phys): Nine physicians independently read the original ultrasound images and make diagnoses.

Complete AI diagnosis (Complete-AI): Use the Explainer and its backbone model (DenseNet-121) to diagnose alone and evaluate their performances.

The Explainer for physicians (Explainer-Phys): Providing the heatmaps and predictions generated by the Explainer, physicians

Table 2
The annotation result for ACR TI-RADS features. Bn: benign. Mal: malignant. N/A: not available. Because the score was left blank when no majority result was reached, the total number of nodules under each feature may not sum to 7714.

Score		0	1	2	3	Total
Composition	Bn	419(9.6%)	585(13.3%)	3382 (77.1%)	N/A	4386
	Mal	11(0.3%)	22(0.7%)	3189(99.0%)	N/A	3222
Echogenicity	Bn	429 (10.6%)	1335(32.9%)	2162 (53.3%)	129 (0.3%)	4055
	Mal	9 (0.03%)	111 (3.9%)	2334 (79.2%)	393 (13.8%)	2847
Shape	Bn	3626 (84.0%)	N/A	N/A	690 (16.0%)	4316
	Mal	1784 (58.7%)	N/A	N/A	1256(41.3%)	3161
Margin	Bn	3312 (75.5%)	1077 (24.5%)	N/A	N/A	4389
	Mal	1535 (50.5%)	1507 (49.5%)	N/A	N/A	3042
Echogenic foci	Bn	3498 (81.3%)	240 (5.6%)	89 (2.1%)	478 (11.0%)	4305
	Mal	1756 (59.1%)	131 (4.4%)	45 (1.5%)	1037(35.0%)	2969

Table 3
The data distribution for each classification task.

Set		Train	Validation	Test	Total
Malignancy	Benign	7968	1149	3826	12943
	Malignant	4205	690	1503	6398
	Total	12173	1839	5329	**19341**
Composition	0	711	93	316	1120
	1	1063	166	506	1735
	2	10209	1544	4421	16174
	Total	11983	1803	5243	**19029**
Echogenicity	0	729	110	310	1149
	1	2378	321	1122	3821
	2	6874	1033	2878	10785
	3	725	115	341	1181
	Total	10706	1579	4651	**16936**
Margin	Normal	7478	1129	3359	11966
	Abnormal	4343	653	1816	6812
	Total	11821	1782	5175	**18778**
Echogenic foci	0	8332	1235	3807	13374
	1	566	122	231	919
	2	200	35	95	330
	3	2386	317	903	3606
	Total	11484	1709	5036	**18229**

were asked to decide whether to accept an AI's forecast based on the heatmap and essential features of the original image in order to assess the benefit of applying the Explainer in a clinical workflow.

The Explainer for TI-RADS features (Explainer-features): Use the Explainer and its backbone (DenseNet-121) to classify four TI-RADS features (margin, composition, echogenicity and echogenic foci) alone and evaluate their performances separately.

3.4. Statistical evaluation

In this paper, we ran 3 trials to evaluate the performance of the Explainer:

a) Results of the Explainer and its backbone model (DenseNet-121) are compared, using the area under receiver operating characteristic (AUROC), accuracy, sensitivity and specificity to evaluate their performance and consistency on a benign-malignant classification task.
b) ‘Complete-Phys’ and‘Explainer-Phys’ compare, using the AUROC, area under precision-recall curve (AUPR), accuracy, sensitivity, specificity, precision, recall and F1-Score to assess whether physicians perform better with the Explainer than diagnose alone.
c) Results of four TI-RADS feature classification tasks of the Explainer and its backbone (DenseNet-121) are compared, using the AUROC, accuracy, sensitivity, specificity to generally evaluate their performance and consistency.

Table 4
Performance and consistency of the Explainer and the DenseNet-121 on malignancy classification task.

	AUROC	Accuracy	Sensitivity	Specificity
DenseNet vs GT	0.887, [0.877,0.897]	82.42%	77.64%	84.29%
Explainer vs GT	0.882, [0.872,0.893]	83.04%	75.58%	85.96%
Explainer vs DenseNet	0.985, [0.983,0.988]	93.02%	93.78%	92.64%

4. Result

4.1. Performance of the Explainer on malignancy classification tasks (Complete-Explainer)

We utilize the Explainer in this part to conduct malignancy classification tasks on our thyroid ultrasound dataset. We proved that the Explainer's performance is equivalent to that of the backbone network (DenseNet-121).

Since the fact that the machine learns the parameters used in the Explainer's map generation process, the Explainer and its backbone network (DenseNet-121) may generate different predictions for the same input image. As a result, we used the DenseNet-121′s AUROC on GT annotation as a benchmark and computed the Explainer's AUROC on DenseNet-121′s prediction to observe how much they vary. The Explainer's performance in the malignancy classification test is summarized in Table 4. We discover no statistically significant differences between the DenseNet-121 and the Explainer; their AUROCs on GT differ by less than 0.005, with mainly overlap 95 percent CI. In addition, a deep ROC analysis [32] is used for measuring the Explainer's AUROC, average sensitivity, and average specificity among groups of different predicted risks as summarized in Table 5.

Table 5
Performance of the Explainer on malignancy classification task in 3 even groups by false positive rate (FPR) for high, medium and low predicted malignancy.

FPR Pred. Risk	[0, 0.33] High	[0.33, 0.67] Medium	[0.67, 1] Low
AUC_{ni}	87.0%	91.0%	97.8%
$\overline{se}_i$	72.5%	96.5%	99.9%
$\overline{sp}_i$	92.1%	45.5%	29.0%

4.2. Results of the multiple reader study

We expect the Explainer to show better ability for explaining which region of its input does the model depend on to predict, and better assist physicians in diagnosis than Grad-CAM. We chose 200 ultrasound images of distinct nodules from the test dataset (Benign:100, Malignant:100) to perform a multiple reader research. Before the AI-assisted reader study, we obtained baseline performances from physicians by asking them to diagnose on their own. Then, we demonstrated the explanation heatmaps created by the Explainer, as well as its malignancy prediction, to aid physicians' diagnosis. Fig. 5 shows the Explanation heatmap functions.

The results of the multiple reader research are summarized in Table 6. The task of reading the original ultrasound images indicates that without the assistance of AI, the F1-Score of the diagnosis of median-level and senior physicians is often greater than the junior physicians. The benefit of using AI assisting physician is obvious, as we can see from Table 6, all physicians had a higher accuracy when reading images with the assistance of the Explainer than without.

To further investigate, we plot the receiver operating characteristics (ROC) and precision-recall (PR) for all physicians, as well as the ROC and PR curves for Explainer. As seen in Fig. 6, only two physicians reached the Explainer's ROC and PR curves by reading the original image, implying that AI is capable of diagnosing on an equivalent or even superior level than humans. Multiple physicians beat AI when assisted by the Explainer, implying that a combination of AI and human physicians would maximize the diagnosis accuracy. Physicians achieved increased precision and recall in all groups when assisted by the Explainer. The trial comparing humans and machines confirms the hypothesis that physicians perform better when assisted by the Explainer.

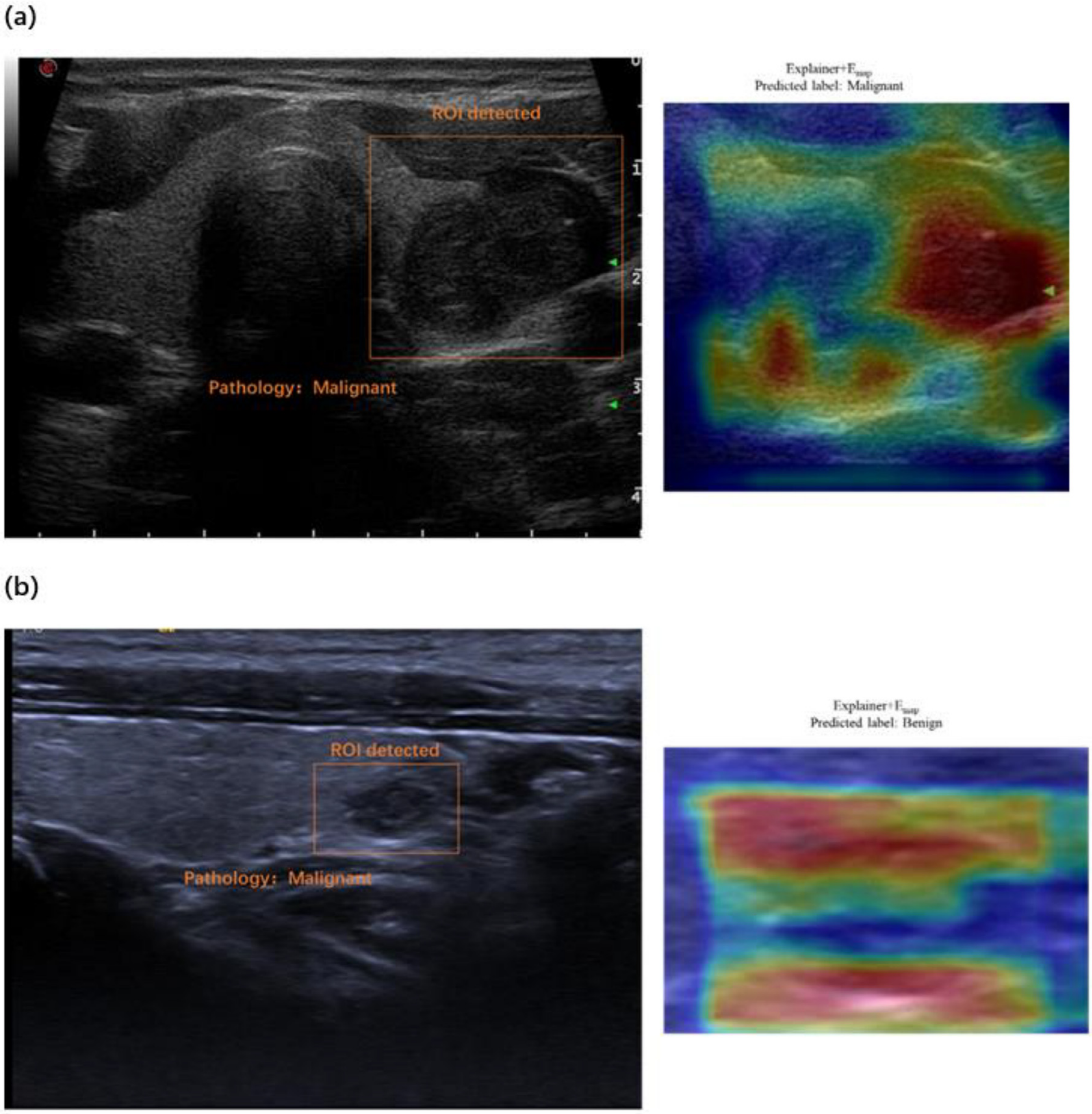
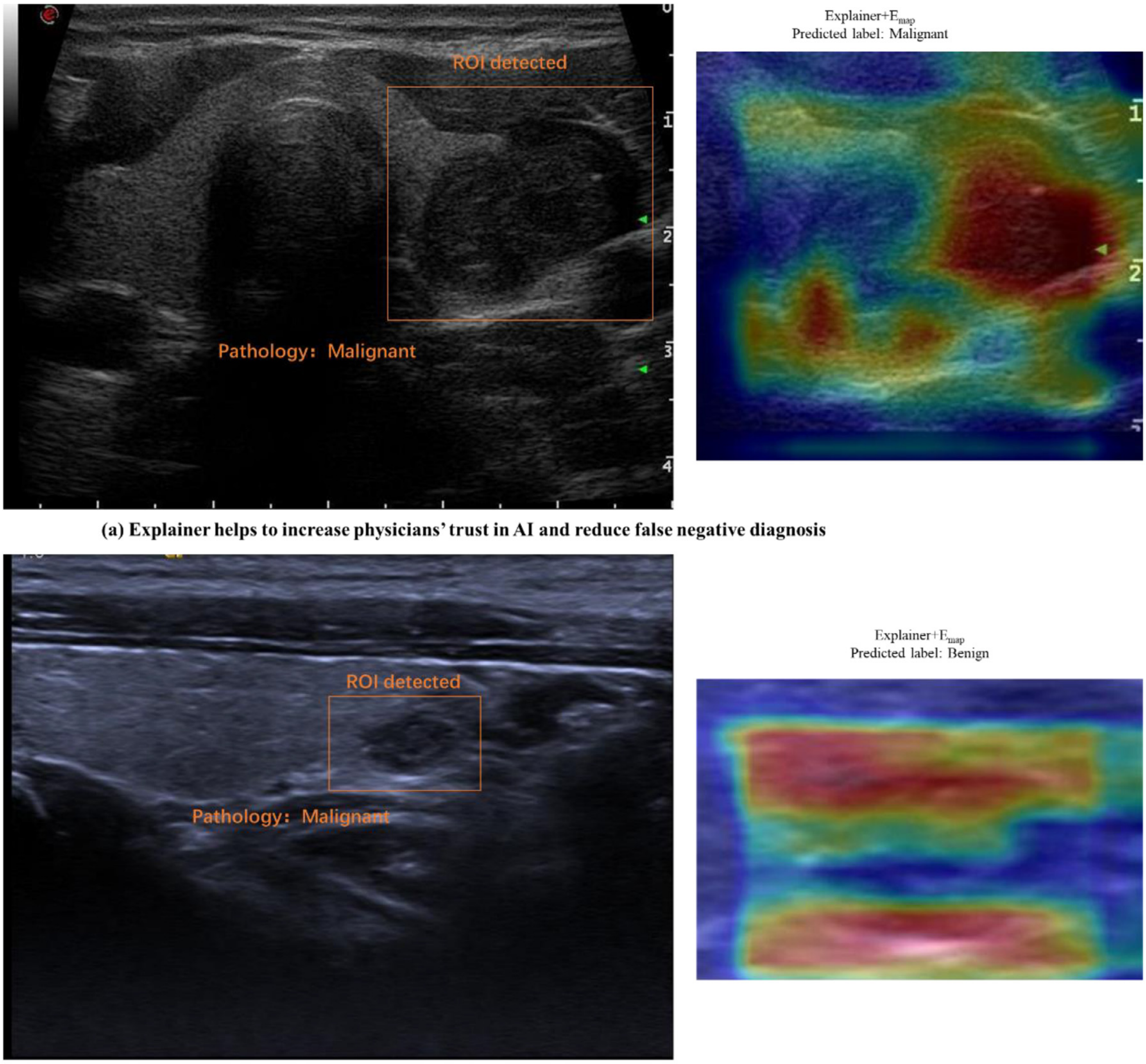


Fig. 5. Two examples from the reader study to illustrate how the explanation heatmap can assist physicians.
(a) The Explainer's heatmap highlights the important tissues near nodule, enhancing physician's faith in AI. (b) Another way to employ heatmaps. We averted a situation where AI may misled physicians; when AI presents an explanation heatmap on the wrong place, physicians may ignore AI's prediction.



(a) Explainer helps to increase physicians' trust in AI and reduce false negative diagnosis

(b) Explainer helps to exclude false-negative case of AI's prediction

Fig. 5. Continued

Table 6
Performance comparation when physicians diagnose alone and assisted by the Explainer.

	Complete Physician					Physician + Explainer				
	Accuracy	Precision	Recall (Sensitivity)	Specificity	F1-Score	Accuracy	Precision	Recall (Sensitivity)	Specificity	F1-Score
Junior-1	75.5%	74.3%	78.0%	73.0%	0.761	79.5%	77.6%	83.0%	76.0%	0.802
Junior -2	65.5%	60.1%	92.0%	39.0%	0.727	72.5%	66.0%	93.0%	52.0%	0.772
Medium-1	79.5%	83.9%	73.0%	86.0%	0.781	82.0%	84.8%	78.0%	86.0%	0.813
Medium-2	78.5%	75.7%	84.0%	73.0%	0.796	85.0%	84.3%	86.0%	84.0%	0.851
Senior-1	78.5%	82.0%	73.0%	84.0%	0.772	81.0%	81.0%	81.0%	81.0%	0.810
Senior-2	75.5%	78.0%	71.0%	80.0%	0.743	79.5%	81.7%	76.0%	83.0%	0.788

4.3. The Explainer vs the DenseNet-121 on ACR TI-RADS-feature classification

To further confirm the consistency between the Explainer and its backbone classifier, we tested them on four ACR TI-RADS-feature classification tasks, including composition, echogenicity, echogenic foci and margin. The data distribution shows that specific types of some TI-RADS characteristics account to the majority. For example, as shown in Table 3, the quantity of lesions with a type 2 composition is approximately ten times as abundant as type 0 and 1, which means that when calculating the precision of type 0, even with a less than 0.1 misclassified type 2 as type 0, and all properly recognized type 0, the precision of type 0 will be less than 0.5. Precision is not appropriate for evaluating the consistency of our thyroid dataset due to high data imbalance. Therefore, we anticipated both models to perform similarly on classification tasks in terms of AUROC, sensitivity, and specificity, and we also computed the Explainer's performance on the DenseNet-121′s prediction to see how much they differ.

Table 7 summarizes the Explainer's classification performance in each task. For all tasks, we find no significant differences between the DenseNet-121 and the Explainer; in all tasks their AU-

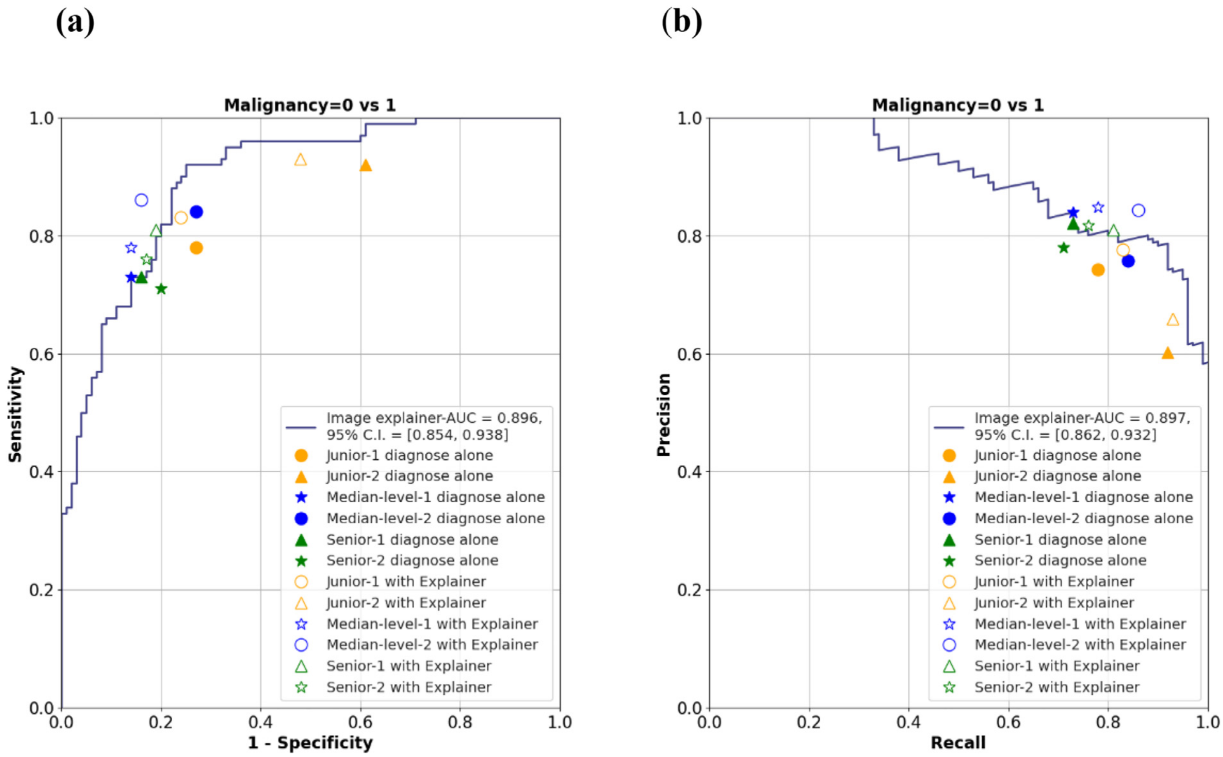


Fig. 6. Receiver operating characteristic (ROC) and precision-recall (PR) comparison of physicians' performance when reading images alone and assisted by the Explainer. (a) The receiver operating characteristics (ROC) for all physicians and the Explainer; (b) The precision-recall (PR) for all physicians and the Explainer. Only two physicians achieved the Explainer's ROC and PR curves. Four professionals outperformed AI when assisted by the Explainer. The Explainer improved precision and recall in all groups.

ROCs on GT differ by less than 0.01, and the biggest differ of all tasks is 0.02 where the DenseNet-121 is slightly better at determining if the nodule's composition corresponds to a score of 0, but the difference is within the 95% confidence interval. When identifying certain types of TI-RADS features, the AUROC of the Explainer on the DenseNet-121's prediction is more than 0.99, and in all types, the number is greater than 0.95.

Meanwhile, we compare the heatmaps generated by the Explainer and the Grad-CAM in all four feature classification tasks. As shown in Fig. 7, the Explainer maps visually locate key regions more accurately. We conclude that predictions made by the Explainer has high consistency with its backbone, and that the Explainer demonstrates a good explanation ability.

Visually, the Explainer maps locate key regions more accurately than Grad-CAM.

5. Discussion

Artificial intelligence has long been acknowledged as a valuable tool in medical diagnosis, and several studies have demonstrated that AI can produce diagnoses that are on par with or better than those made by physicians [33–35]. But the black box of AI's prediction process presents physicians with a perplexing dilemma when deciding whether to believe its diagnosis, particularly when the physician's initial diagnosis contradicts AI's forecast. Trusted AI solutions must be able to cope with imprecision, missing, and incorrect information and explain both the result and the process of how it was obtained to a medical expert [25]. To improve the interpretability of the thyroid AI system in our study, we propose a novel framework called Explainer, which is a tiny CNN parasitic network capable of convolutionally constructing an E_{map} to explain the CNN prediction dependent region. To test the performance of the Explainer, we apply it to different classification tasks, including benign-malignant classification and TI-RADS feature classification (margin, composition, echogenicity, and echogenic foci). Results show that the explanation heatmaps generated by the Explainer locate a more reasonable region of thyroid lesion than Grad-CAM. Furthermore, multi-center multi-readers study of thyroid benign-malignant classification reveals that physicians perform better when assisted by the Explainer.

In our study, the Explainer is an intrinsic model specific XAI for thyroid classification. The Explainer uses CNN to generate feature-level masks and occlude portions of feature maps, where only the information of portions not occluded will be sent to a classifier. By mapping it to the original input image, it highlights the regions that the model uses to make predictions.

The Explainer is a tool that assists physicians in determining if AI is spotting the proper location and discovering some lesion characteristics that the physician may overlook by describing AI's interested region. If AI focuses on specific parts of the lesion and peripheral tissues, physicians could feel more secure in accepting AI's diagnosis. If AI focuses on parts that are remote from the lesion and lacks important traits, physicians may need to reject its predication and re-evaluate the case.

Using two typical examples from our reader research results, our study demonstrates how AI assists physicians in diagnosing and why an explanation heatmap is critical in this process. In Fig. 5(a), only two out of six physicians accurately diagnose the malignancy when it is diagnosed alone. By providing the Explainer's heatmap, which comprises most of the nodule's key tissues, it significantly increases physicians' trust in AI and allows them to reevaluate the nodule's malignancy. Fig. 5(b) demonstrates another alternative approach to using heatmaps; the Explainer gave an incorrect prediction for this nodule, which is also in contrast to physicians' self-prediction. In this scenario, we avoided the circumstance in which the AI misled physicians by providing ex-

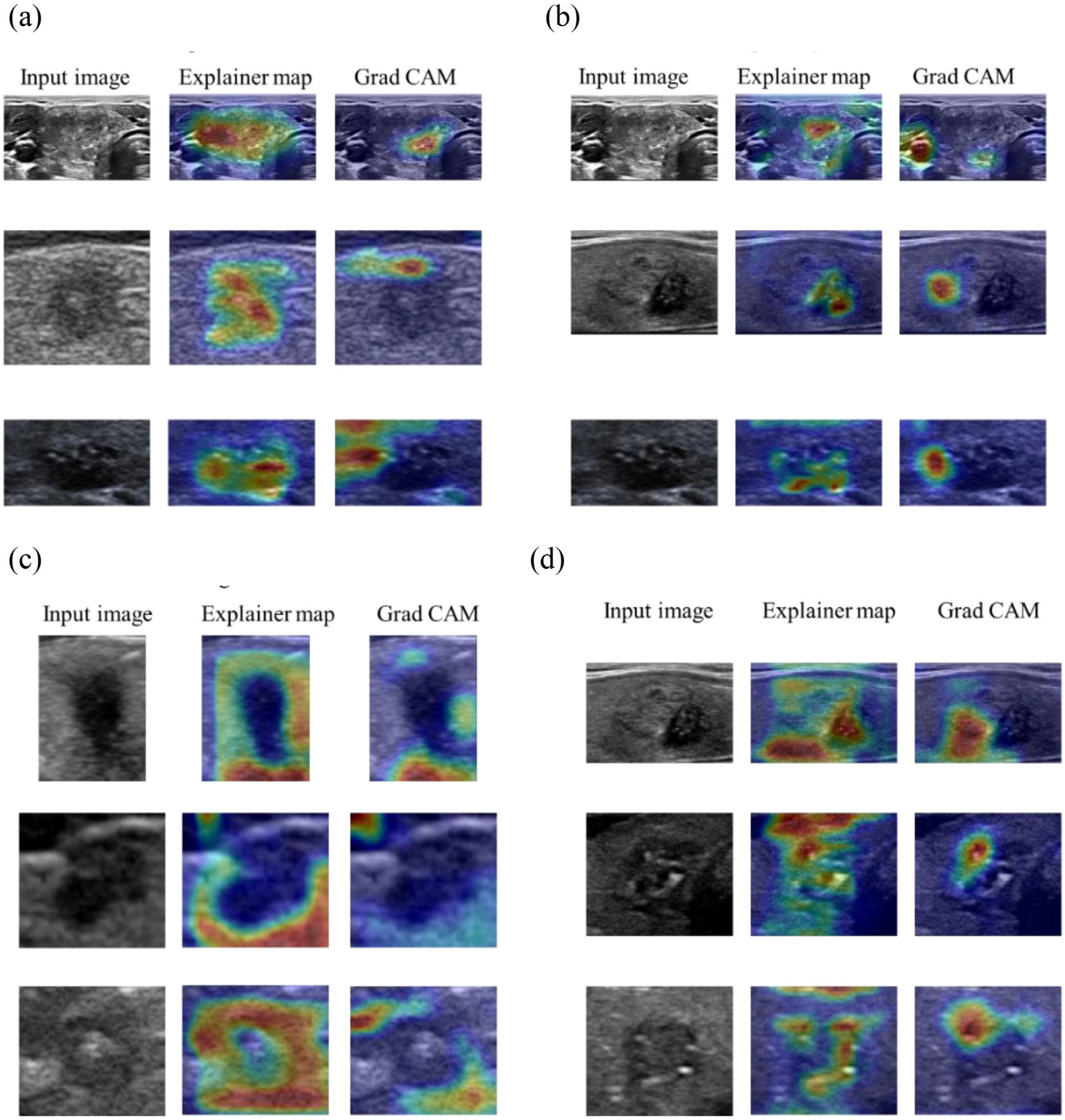


Fig. 7. Heatmaps of Explainer and Grad-CAM for four feature classification tasks (a) Task: Composition classification; (b) Task: Echogenicity classification; (c) Task: Margin classification; (d) Task: Echogenic foci classification.

planatory heatmaps indicating that the AI was focusing on the incorrect location and that its prediction could not be believed.

According to a multiple reader study, when the explainer is used to aid in diagnosis, the physician performs better than when they diagnose alone. Physicians who use an explainer perform better in terms of precision and recall. The Explainer improved diagnostic accuracy and efficiency by streamlining the thyroid diagnosis workflow and directing physicians to focus on important features, particularly for less experienced junior physicians.

The Explainer's utility isn't limited to thyroid disorders, and we're looking into expanding its use in the Ultrasound Breast Cancer Diagnosis Mission. According to the ACR BI-RADS Atlas, 5th Edition [36], more than 20 features are used in diagnosing breast cancer, including different kinds of margin, echo pattern, composition, and so on. Integrating the Explainer into the CNN-based BI-RADS feature detector and malignancy predictor would certainly improve the risk assessment of breast disorders. Besides, we are also looking forward to applying the Explainer to other formats of data, such as videos, which are widely used in ultrasound tests.

Our study has several limitations. A disease's diagnostic dimensions can be multiple, resulting in numerous patient subgroups. Because the quantity and quality of data utilized in deep learning are so critical, the chase for the optimal high-volume dataset is never-ending. Meanwhile, the incidence of many disease subtypes is very rare. As we have mentioned before, the distribution of the TI-RADS features in our current dataset is extremely imbalanced. To achieve a better diagnostic AI and test its generalizability, we expect to involve more hospitals and physician groups in multi-center clinical trials.

Table 7
Performance of the Explainer and the DenseNet-121 on ACR TI-RADS classification tasks.

		0 vs other			1 vs other			2 vs other			3 vs other		
		DenseNet vs GT	Explainer vs GT	Explainer vs DenseNet	DenseNet vs GT	Explainer vs GT	Explainer vs DenseNet	DenseNet vs GT	Explainer vs GT	Explainer vs DenseNet	DenseNet vs GT	Explainer vs GT	Explainer vs DenseNet
Composition	AUROC	0.861, [0.837,0.885]	0.840, [0.812,0.867]	0.996, [0.995,0.998]	0.839, [0.819,0.858]	0.830, [0.809,0.851]	0.994, [0.992,0.995]	0.907, [0.895,0.919]	0.903, [0.890,0.916]	0.993, [0.992,0.995]	N/A	N/A	N/A
	Sensetivity	74.68%	70.89%	98.37%	69.37%	70.16%	95.78%	83.17%	84.98%	95.67%	N/A	N/A	N/A
	Specificity	84.17%	84.29%	96.36%	85.56%	85.60%	96.76%	84.31%	83.21%	96.61%	N/A	N/A	N/A
Echogenic foci	AUROC	0.865, [0.852,0.877]	0.867, [0.854,0.879]	0.994, [0.993,0.996]	0.796, [0.762,0.830]	0.788, [0.754,0.822]	0.996, [0.994,0.997]	0.866, [0.826,0.907]	0.866, [0.825,0.907]	0.999, [0.998,1.000]	0.831, [0.817,0.846]	0.838, [0.823,0.852]	0.985, [0.982,0.988]
	Sensetivity	82.85%	78.20%	94.54%	68.40%	73.16%	99.51%	74.12%	68.24%	94.12%	74.53%	73.53%	94.72%
	Specificity	74.98%	80.39%	98.07%	79.87%	72.95%	95.27%	83.04%	90.24%	99.48%	77.61%	80.77%	93.24%
Echogenicity	AUROC	0.883, [0.862,0.904]	0.882, [0.861,0.903]	0.987, [0.984,0.990]	0.817, [0.803,0.831]	0.819, [0.805,0.833]	0.958, [0.953,0.963]	0.788, [0.775,0.802]	0.786, [0.773,0.800]	0.957, [0.953,0.963]	0.798, [0.774,0.821]	0.798, [0.775,0.821]	0.990, [0.986,0.993]
	Sensetivity	78.39%	78.71%	96.77%	74.42%	73.98%	88.94%	75.36%	74.22%	88.94%	74.78%	76.25%	97.33%
	Specificity	86.69%	85.05%	93.83%	74.61%	75.69%	88.01%	68.47%	69.88%	88.01%	69.40%	68.72%	95.94%
Margin	AUROC	0.740, [0.727,0.754]	0.745, [0.732,0.758]	0.975, [0.971,0.979]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Sensetivity	63.17%	55.25%	91.29%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Specificity	72.69%	81.88%	90.02%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

6. Conclusion

To summarize, we developed a system called Explainer, which is capable of classifying thyroid ultrasound images while generating a heatmap to explain its prediction. Reader studies show that ultrasound specialists perform better diagnostically when assisted by the Explainer than when diagnosing alone. It provides physicians with a tool for assessing the reliability of AI diagnosis, particularly when their initial diagnosis is opposed to AI. As a result, incorporating Explainer into the clinical ultrasound screening workflow may provide real benefits by enabling physicians to make more accurate diagnoses.

CRediT authorship contribution statement

Fajin Dong, Conceived and designed the experiments, Analyzed the data, Wrote the paper.

Di Song, Conceived and designed the experiments, Analyzed the data, Wrote the paper.

Jincao Yao, Conceived and designed the experiments, Analyzed the data, Wrote the paper.

Jinfeng Xu, Conceived and designed the experiments, Analyzed the data, Wrote the paper.

Siyuan Shi, Performed the experiments, Analyzed the data.

Chen Cui, Performed the experiments, Analyzed the data.

Yitao Jiang, Performed the experiments, Analyzed the data.

Liping Wang, Performed the experiments, Analyzed the data.

Huaiyu Wu, Analyzed the data.

Lijing Wang, Analyzed the data.

Hongtian Tian, Analyzed the data.

Di Ou, Analyzed the data.

Linghu Wu, Analyzed the data.

Wei Li, Analyzed the data.

Xiuqin Ye, Analyzed the data.

Na Feng, Analyzed the data.

Weiyun Pan, Analyzed the data.

Mei Song, Analyzed the data.

Dong Xu, Conceived and designed the experiments, Analyzed the data, Contributed materials/analysis tools.

Declaration of Competing Interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

Data availability

The raw data generated in this study is not publicly available due to the need for patient privacy protection but is available upon reasonable request from the corresponding author.

Financial support

No.

Code availability

The code for the thyroid explainer model framework is provided for reproduction. The code was written in Python v3.9. It is publicly available at https://data.mendeley.com/datasets/gg24hxmj7v.

Acknowledgments

This project was supported by Commission of Science and Technology of Shenzhen (GJHZ20200731095401004).

References

[1] S Vaccarella, L. Dal Maso, Challenges in investigating risk factors for thyroid cancer, The Lancet Diabetes&Endocrinology 9 (2) (2021) 57–59.

[2] A Miranda-Filho, J Lortet-Tieulent, F Bray, B Cao, S Franceschi, S Vaccarella, et al., Thyroid cancer incidence trends by histology in 25 countries: a population-based study, The Lancet Diabetes&Endocrinology 9 (4) (2021) 225–234.

[3] U Raghavendra, UR Acharya, A Gudigar, JH Tan, H Fujita, Y Hagiwara, et al., Fusion of spatial gray level dependency and fractal texture features for the characterization of thyroid lesions, Ultrasonics 77 (2017) 110–120.

[4] CM Kitahara, FD Körmendiné, JOL Jørgensen, D Cronin-Fenton, T. SørensenHenrik, Benign thyroid diseases and risk of thyroid cancer: a nationwide cohort study, The Journal of Clinical Endocrinology Metabolism 103 (6) (2018) 2216–2224.

[5] S Kang, E Lee, CW Chung, HN Jang, JH Moon, Y Shin, et al., A Beneficial Role of Computer-aided Diagnosis System for Less Experienced Physicians in the Diagnosis of, Thyroid Nodule on Ultrasound. Scientific Reports. 11 (1) (2021) 20448.

[6] R Yang, X Zou, H Zeng, Y Zhao, X. Ma, Comparison of diagnostic performance of five different ultrasound TI-RADS classification guidelines for thyroid nodules, Frontiers in Oncology (2020) 2457.

[7] M Buda, B Wildman-Tobriner, JK Hoang, D Thayer, FN Tessler, WD Middleton, et al., Management of thyroid nodules seen on US images: deep learning may match performance of radiologists, Radiology 292 (3) (2019) 695–701.

[8] T Liu, Q Guo, C Lian, X Ren, S Liang, J Yu, et al., Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks, Medical image analysis 58 (2019) 101555.

[9] F Verburg, C. Reiners, Sonographic diagnosis of thyroid cancer with support of AI, Nature Reviews Endocrinology 15 (6) (2019) 319–321.

[10] X Wei, J Zhu, H Zhang, H Gao, R Yu, Z Liu, et al., Visual interpretability in computer-assisted diagnosis of thyroid nodules using ultrasound images, Medical science monitor: international medical journal of experimentalclinical research 26 (2020) e927007-1.

[11] G Yang, Q Ye, J. Xia, Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond, Information Fusion 77 (2022) 29–52.

[12] P Croskerry, K Cosby, ML Graber, H. Singh, Diagnosis: Interpreting the shadows, CRC Press, 2017.

[13] K-H Yu, AL Beam, Kohane IS. Artificial intelligence in healthcare, Nature biomedical engineering 2 (10) (2018) 719–731.

[14] S Hajian, F Bonchi, C Castillo, editors. Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. the 22nd ACM SIGKDD International Conference; 2016.

[15] H Liu, Y Wang, W Fan, X Liu, Y Li, S Jain, et al. Trustworthy AI: A Computational Perspective. 2021.

[16] A Holzinger, The Next Frontier: AI We Can Really Trust, CCIS 2021 (1524) 427–440.

[17] AB Arrieta, N Díaz-Rodríguez, J Del Ser, A Bennetot, S Tabik, A Barbado, et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115.

[18] RR Selvaraju, M Cogswell, A Das, R Vedantam, D Parikh, D Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017.

[19] K Simonyan, A Vedaldi, A Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, Workshop at International Conference on Learning Representations, ICLR, 2014.

[20] MD Zeiler, R Fergus, Visualizing and understanding convolutional networks. European conference on computer vision editors, Springer, 2014.

[21] Z Zhang, Y Xie, F Xing, M McGough, L Yang, Mdnet: A semantically and visually interpretable medical image diagnosis network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

[22] F Dong, R She, C Cui, S Shi, X Hu, J Zeng, et al., One step further into the blackbox: a pilot study of how to build more confidence around an AI-based decision system of breast nodule assessment in 2D ultrasound, European Radiology (2021) 1–10.

[23] A Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE access 6 (2018) 52138–52160.

[24] E Tjoa, C. Guan, A survey on explainable artificial intelligence (xai): Toward medical xai, IEEE Transactions on Neural Networks Learning Systems 32 (11) (2020) 4793–4813.

[25] A Holzinger, M Dehmer, F Emmert-Streib, R Cucchiara, I Augenstein, JD Ser, et al., Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence 79 (2022) 263–278.

[26] T-Y Lin, P Goyal, R Girshick, K He, P Dollor, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017.

[27] G Huang, Z Liu, L Van Der Maaten, KQ Weinberger, editors. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017.

[28] K He, X Zhang, S Ren, J Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[29] F Chollet, Xception editor, Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

[30] AG Howard, M Zhu, B Chen, D Kalenichenko, W Wang, T Weyand, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:04861. 2017.

[31] I Culjak, D Abram, T Pribanic, H Dzapo, M Cifrek, A brief introduction to OpenCV. 2012 proceedings of the 35th international convention MIPRO editors, IEEE, 2012.

[32] AM Carrington, DG Manuel, PW Fieguth, T Ramsay, V Osmani, B Wernly, et al., Deep ROC Analysis and AUC as Balanced Average Accuracy for Improved Classifier Selection, Audit and Explanation, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (1) (2023) 329–341.

[33] Y Shen, FE Shamout, JR Oliver, J Witowski, K Kannan, J Park, et al. Artificial Intelligence System Reduces False-Positive Findings in the Interpretation of Breast Ultrasound Exams. medRxiv. 2021.

[34] L-Q Zhou, X-L Wu, S-Y Huang, G-G Wu, H-R Ye, Q Wei, et al., Lymph node metastasis prediction from primary breast cancer US images using deep learning, Radiology 294 (1) (2020) 19–28.

[35] VL Mango, M Sun, RT Wynn, R. Ha, Should we ignore, follow, or biopsy? Impact of artificial intelligence decision support on breast ultrasound lesion assessment, American Journal of Roentgenology 214 (6) (2020) 1445–1452.

[36] DA Spak, J Plaxco, L Santiago, M Dryden, B. Dogan, BI-RADS® fifth edition: A summary of changes, Diagnostic interventional imaging 98 (3) (2017) 179–190.