IEEE Access
Multidisciplinary ⁞ Rapid Review ⁞ Open Access Journal

Received December 27, 2019, accepted February 17, 2020, date of publication February 27, 2020, date of current version March 26, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2976495

A Novel Interpretable Computer-Aided Diagnosis System of Thyroid Nodules on Ultrasound Based on Clinical Experience

SHIJIE ZHANG[1], HUARUI DU[1], ZHUANG JIN[2], YAQIONG ZHU[2], YING ZHANG[2], FANG XIE[2], MINGBO ZHANG[2], XIAOQI TIAN[2], JUE ZHANG[1,3], AND YUKUN LUO[2]

[1]Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China
[2]Ultrasonography Department, General Hospital of the People's Liberation Army, Beijing 100039, China
[3]College of Engineering, Peking University, Beijing 100871, China

Corresponding authors: Jue Zhang (zhangjue@pku.edu.cn) and Yukun Luo (lyk301@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 81771834, in part by the Natural Science Foundation of Beijing Municipality under Grant 7194318, and in part by the Science and Technology Program of Beijing under Grant Z181100001718017.

ABSTRACT Computer-aided diagnosis systems (CADs) present valuable second opinions to radiologists in diagnosis. Many studies on thyroid nodules have proposed various CADs to provide a binary result, benignity or malignancy, for doctors, ignoring interpretability of more ultrasonic features that could be more useful. We develop an interpretable CADs (iCADS) that utilizes deep-learning networks' classification power and interpretability potential of clinical guidelines, like TIRADS, a well-established scale for thyroid nodules. iCADS incorporates a main neural-networks model and six neural-network based interpreters. The outputs of the six interpreters are compared with TIRADS guidelines and the matched result will form a report, more than a benignity or malignancy result, for radiologists. Clinical images of 16,946 thyroid nodules from 5,885 patients were used to train the proposed iCADS. An extra experimental data set containing 501 images were used to test the performance of the model. For better illustrating the assistant ability of iCADS, we also recruited ten junior radiologists to make diagnosis decisions with or without the help of different versions of iCADS. The experiments demonstrated that iCADS can largely improve junior radiologists diagnosis with the help of interpreter strategy. These experiments are also the very first attempt to evaluate the effect of interpretability of deep-learning based CADs in clinical practice. Comparison experiments with other deep-learning based CADs and traditional CADs indicated that the interpreter strategy can easily be combined to other intelligent CADs without the loss of performance. The framework of iCADS can also inspire more research on the development of CADs.

INDEX TERMS Interpretable computer-aided diagnosis system, deep learning, multi-task learning, thyroid nodules, ultrasound.

I. INTRODUCTION

Thyroid cancer is one of the most serious cancers among endocrine tumors, and its incidence has been rising more rapidly than other types of cancers [1]. High-resolution ultrasound, thanks to its lower cost, noninvasive scanning and non-radiation, has been widely used to image thyroid nodules. But distinguishing malignant nodules from benign ones through ultrasound images is heavily dependent on the experience of radiologists, and a junior physician can easily misinterpret those sonographic characteristics [2], [3].

Computer-aided diagnosis systems (CADS) can provide efficient and quantitative diagnostic results to assist doctors in the interpretation of medical images. A CAD result can be seen as a second opinion that is objective and comprehensive. Studies have reported that the second opinion provided by CADS, known as a double reading, contributed to the reduction of the occurrence of missed cancers [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Luca Cassano.

Traditional approaches utilized features that were manually extracted from thyroid ultrasound images [5]–[7]. However, Liu T et al found that these features were inept for their dataset [8] and proposed a new extraction method that combines deep features with conventional textual findings.

This improvement, however, can hardly cope with the style of varies images. Given different ultrasonic devices and diversified radiologists' scanning habits, ultrasonic images can differ widely from dataset to dataset. A high-transferability classification model is required.

Some researchers have adopted deep learning to classify diseases features through ultrasonic images, including breast lesions [9], liver lesions [10], and prostate lesions [11].

These deep-learning-based approaches can indeed compete with skilled radiologists [7], [12]–[14]. but few studies have regarded their expertise as an assistant to doctors. It has been widely acknowledged that the goal of CADs is not to replace radiologists, but to deliver valuable second opinions to them [15].

Actually, deep-learning algorithms are far from being as trustworthy as people think. Some researchers have even claimed" letting the data speak for themselves can be problematic" [16]. Machine learning could present some misclassifications simply against human's common sense [17], [18]. These errors could come from parameter sampling, or certain noises in the original images. Given possible medical malpractice these mistakes might cause, AI diagnosis needs strict double-check from physicians. Their contribution is to provide intelligent references to doctors, not to act for them.

But the existing classification models output results in terms of statistical texture features or deep network features that doctors can hardly understand. Although some literatures claimed their approaches could elevate radiologists' ability to diagnose thyroid cancer [12], they failed to elaborate how much the CADs contributed to final decisions.

Both traditional and deep-learning approaches output a binary result: benign or malignant. The lack of inference pathway cannot convince radiologists [15], and CADs that could provide more interpretable and elaborative results are eagerly required.

In fact, many clinical guidelines can be used as reference to CADs, especially the medical-image-related reporting systems. For example, in 2009, an ultrasonogram reporting system for thyroid nodules proposed by Horvath *et al.* [19]. In 2009, guidelines for the management of patients with thyroid nodules and differentiated thyroid cancers developed by the American Thyroid Association (ATA) in 2015 [20]. ACR TIRADS Committee's recommendations [21] were released by the ACR Thyroid Imaging, Reporting and Data System (TIRADS) Committee in 2017. These clinical guidelines hold compelling potentials to retrofit deep-learning-based CADs.

Among them, ACR TIRADS is the dominant guidelines for determination on thyroid nodules. It is a semi-quantitative scale that describes ultrasound features designating to benign or malignant nodules and assigns scores for each level. Thyroid nodules are divided into 5 levels where a larger number implies a higher risk of malignancy. Some studies found that based on consent of multi-doctors on the TIRADS-based results, the classification could reach a correctness rate of 80.9% [22]. Similarly, some doctors input BIRADS(guidelines for breast nodules)-based features into classifiers and the correctness passed 90% [23]. Fig. 1 displays an ultrasound image of a thyroid nodule. According to ACR TIRADS, its four features beget it 7 points: solid or almost completely solid(2 points), Hypoechoic(2 points), wider-than-tall(3 points), smooth(0 point), which leads to highly suspicious malignancy. And the pathological result confirmed that.

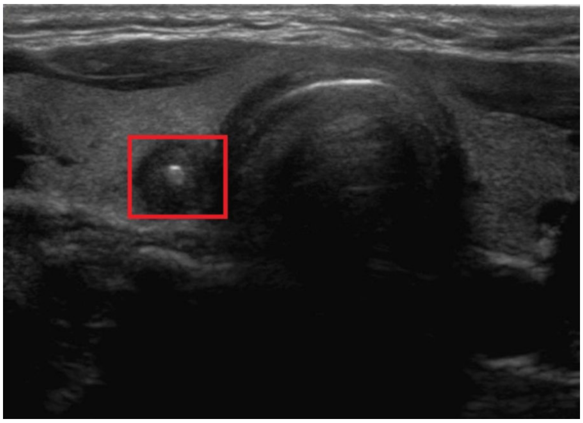


FIGURE 1. Typical ultrasound image of thyroid nodule.

Much different from common strategies that are most likely to improve the accuracy of classifiers in previous CADs, the starting point of this study is to develop an interpretation computer aided diagnosis system (iCADS) to achieve better decision assistant's ability by introducing clinical guidelines.

A common deep learning based CADs can effectively explore the underlying patterns among ultrasound images and the corresponding pathology results through the supervised learning process, which can achieve highly-accurate but seem difficult to interpret. Basically, one feasible solution for this issue is to ignore the pathology results, and develop a series of clinical feature classifiers to learn from radiologist's experiences. However, the performance are limited by the experiences of radiologists, and probably cannot compete to the pathology learning based CADs

In this study, we appropriate the classification power of a deep-learning algorithm, ResNet50, and the valuable referencing of ACR TIRADS to establish a novel iCADS for ultrasound thyroid nodule images. Our system extends normal ResNet50 into a framework containing seven classifiers: one for judging benignity or malignancy, and six for extract nodules features. Through matching the features with ACR TIRADS, iCADS can provides understandable reference to radiologists. The comtributions of our study are as follows.

1. The extended six classifiers can exploit deep features that convolutional network extracted and transforms them

into comprehensive information with the help of TIRADS. Thus, iCAD can provides valuable reference to doctors, more than just benignity or malignancy, which elevates the assistant ability of CADs into a higher level.

2. The extra information iCADS interprets can improve diagnosis correctness of junior radiologists, which was validated by a series of experiments. And these experiments are the first ones that could evaluate the effect of interpretability of CADs in clinical practice.

II. METHOD

A. DATA COLLECTION

Our retrospective study was approved by the Research Ethics Committee of the General Hospital of the People's Liberation Army of China. The two datasets used in this study are the training dataset for model training and validation, and the experimental dataset for evaluation. No image of the same patient is included in both datasets. The ultrasonic scanners that obtained all the images of the two datasets included Siemens ACUSON S2000, Philips iU22, Esaote MyLab Twice, the Siemens ACUSON SEQUOIA 512, the Hitachi HI VISION Ascendus, PHILIPS iU Elite and GE Vivid E9 ultrasound systems with a high-frequency probe.

The training data set contains 16,946 images collected from 5,885 patients with thyroid nodules between May 2014 and November 2018.

An extra experimental data set includes 501 images from 300 patients with thyroid nodules from June 2016 to June 2017.

Images in both datasets had been labeled with pathological tags before they were transferred to investigators, and 4,078 of the training set and all the images in the experimental set were marked with TIRADS by an experienced radiologist.

The inclusion criteria are (1) patients with complete preoperative ultrasound of the thyroid nodules, (2) patients who underwent surgery or a core needle biopsy(CNB) after thyroid examination, (3) patients who underwent a fine needle aspiration biopsy (FNAB) for thyroid benign lesions (excluding adenomas) at least two times with a one-year interval, (4) patients who underwent initial FNAB and US follow-up (>12 months after FNAB) for thyroid benign lesions (excluding adenomas). Eligibility criteria are specified as: patients were excluded if pathological findings were inflammatory lesions or unclear.

B. INTERPRETABLE AIDED DIAGNOSIS SYSTEM

The proposed iCADS is to utilize intelligent decisions of the deep learning model as well as the informative features it extracts. With the help of clinical guidelines, the extracted features can be transformed into comprehensive information and used as reference, together with the result of the deep-learning model, to facilitate radiologists' final decision.

iCADS consists of two parts: a main network trained by pathological images to output benign or malignant, and an interpretation strategy containing six interpreter networks based on clinical guidelines to explain the incomprehensible features extracted by the main network.

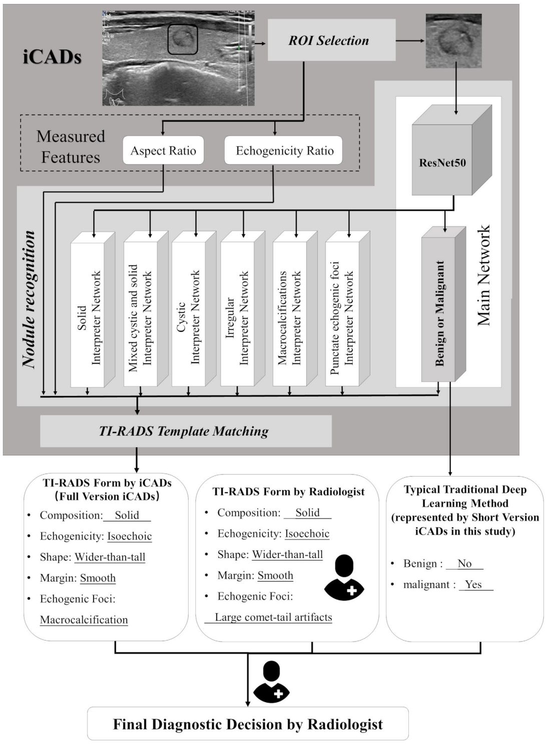

iCADs
ROI Selection
Measured Features
Aspect Ratio
Echogenicity Ratio
ResNet50
Nodule recognition
Solid Interpreter Network
Mixed cystic and solid Interpreter Network
Cystic Interpreter Network
Irregular Interpreter Network
Macrocalcifications Interpreter Network
Punctate echogenic foci Interpreter Network
Benign or Malignant
Main Network
TI-RADS Template Matching
TI-RADS Form by iCADs (Full Version iCADs)
Composition: Solid
Echogenicity: Isoechoic
Shape: Wider-than-tall
Margin: Smooth
Echogenic Foci: Macrocalcification
TI-RADS Form by Radiologist
Composition: Solid
Echogenicity: Isoechoic
Shape: Wider-than-tall
Margin: Smooth
Echogenic Foci: Large comet-tail artifacts
Typical Traditional Deep Learning Method (represented by Short Version iCADs in this study)
Benign : No
malignant : Yes
Final Diagnostic Decision by Radiologist

FIGURE 2. The diagram of the proposed iCADS structure.

First, iCADS selects region of interest (ROI) of thyroid nodules from original ultrasound images; second, deep networks classify nodules, including main networks and six interpreters. Finally, iCADS compare nine outputs of classifiers with TIRADS and presents detailed and final suggestion to radiologists. In practical terms, the clinical features were identified and formed diagnostic report by the iCADS and one radiologist independently. Thereafter, the radiologist considers both diagnoses and draws the final diagnostic decision based on TIRADS. The three-step processing is elaborated as follows.

1) ROI SELECTION (PREPROCESS)

In this procedure, a single roi covering an entire nodule in each image is manually selected by an experienced radiologist. Each ROI is a square and normalized to a size of 224 × 224. Then the aspect and the echogenicity ratio can be calculated by following equations respectively:

$$R_a spect = \frac{w_{roi}}{h_{roi}} \tag{1}$$

$$R_{echogenicity} = \frac{E_n}{E_t} \tag{2}$$

where w_{roi} is the width of ROI, h_{roi} is the height of ROI, E_n is the average intensity in ROI and E_t is the average intensity around the ROI, as shown in Fig. 3. The two parameters will be used as two of the outputs of iCAD for template matching.

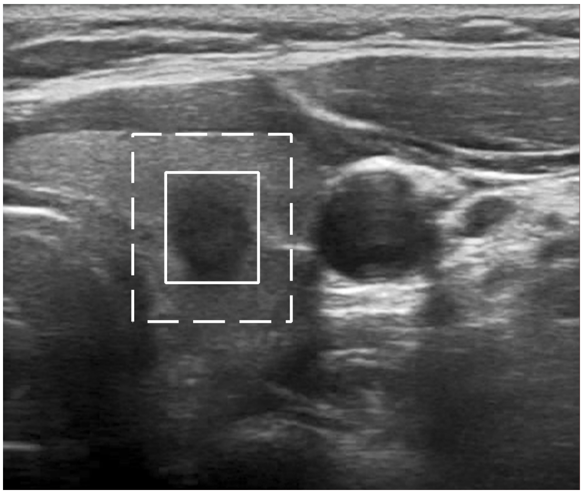


FIGURE 3. Typical ultrasound ROI section of thyroid nodule.

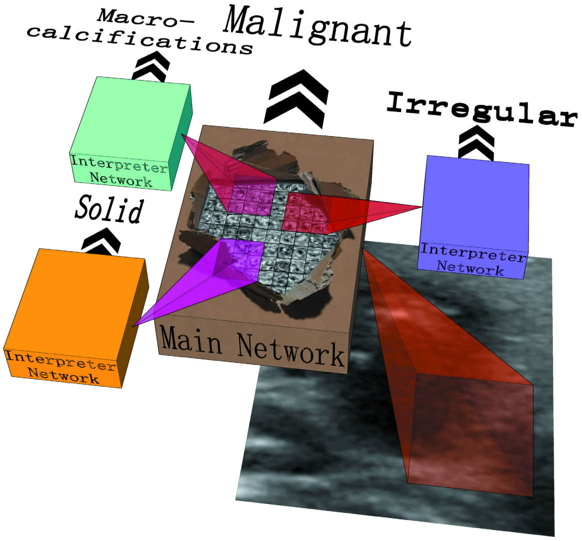


FIGURE 4. Main network and interpreter network.

2) NODULE RECONGNITION

An increasing number of researchers have delved into the interpretability of deep networks, and developed models that can generate interpretable knowledge representations [24]. These approaches inspire us to exploit deep features hiding in the layers of neural networks, but their resulted knowledge representations are still obscure for clinical practice. Therefore, we design an interpretation method that maps deep features into clinical characteristics to obtain more explicit representations.

The whole iCADs is constructed based on the extended ResNet50 networks, and composed of two parts: the main networks to classify nodules into benign or malignant =, ones, and the interpretation model composed of 6 six interpreter networks to classify clinical features, as shown in Fig. 4. It is the deep features extracted by the main networks, instead of the nodule ultrasound images, that are input into the interpretation model.

Fig. 2 indicates that the training procedure is actually a multi-task learning process and its key drive is the ResNet50. The ResNet50 [25] used in this study was pre-trained on ILSVRC [26], which have demonstrated useful for medical image analysis [27]. As a typical DCNN, the Resnet50 ResNet50 consists of five stages of convolutional layers, and an average pooling layer followed by a 1000-way fully connected layer. To tailor to the multi-task strategy, we retrofit the ResNet50's original framework by extending it upon the most commonly used Shared-Bottom multi-task DNN structure [28], that is, replacing the 1000-way fully connected layer with seven parallel separate layers. The seven layers correspond to the respect seven binary classification tasks, including benign or malignant, solid or not, cystic or not, mixed cystic and solid or not, with or without macrocalcifications, with or without punctate echogenic foci, and smooth or irregular, as shown in Table.1. Each task outputs a binary result, which can be easily normalized and fitful for later TIRADS form template matching. Each network training clinical feature has a separate layer with 256-dimensional fully connected layers followed by a two-way softmax layer with randomly initialized weights W drawn from a normal distribution as follows: $W \sim N(\mu = 0, \sigma^2 = 0.05)$. It is worth mentioning that some features mentioned in the ACR TIRADS guide were not included in the model due to its poor robustness, which appeared less in the training data. Besides, features of shape and echogenicity in ROI were estimated by direct measurement, and did not participate in network training.

Notably, the six features that are labeled in the ultrasound images and fed into the six impetrators are seriously unbalanced. A common solution is to sample those data before training. This is suitable for most single task classification, but would cause oversampling or under-sampling in multi-tasks and even exasperate the unbalance. Therefore, we deploy a strategy of rotation training, as explained in Fig.5. During each round, the network is trained feature by feature with each batch; then it is trained for the classifier that predicts benignity and malignancy. This rotation training keeps repeating until all the feature-learning processes reach convergence. All the tasks employ the categorical hinge loss for training with the gradient descent method.

3) TI-RADS FORM TEMPLATE MATCHING

iCADS uses ACR TIRADS to match the networks' outputs with clinical characteristics. ACR TIRADS lists 18 features in an ultrasound image that could imply the state of a thyroid nodule, leading to 262144(218) potential feature combinations. Fortunately, most of these feature combinations are unlikely to occur in clinics. We selected the most representative 13 features, as listed in Table. 1, and form 83 feature combinations for matching templates. Thus, those nine outputs of iCADS, including six features from six interpreters, one benign or malignant classification and two measurement parameters of the ROIs, can be compared with these

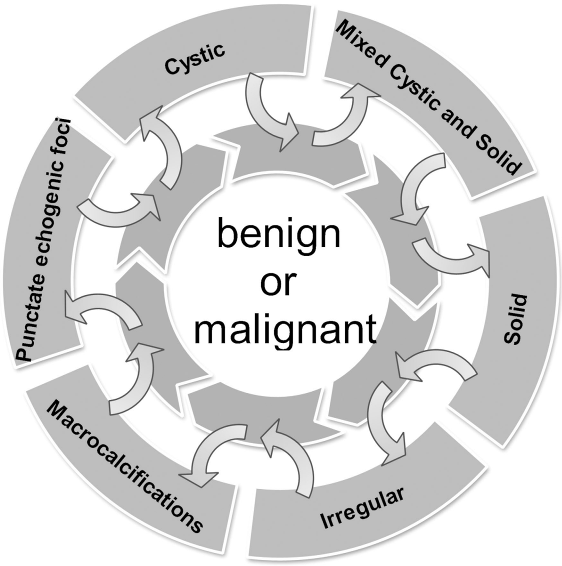


FIGURE 5. Training process is a continuous cycle. Each task is trained alternately.

TABLE 1. The features mentioned in TIRADS and their detection methods in this study.

Categories	Clinical Features	Detector
COMPOSI-TION	Cystic or almost completely cystic	Interpreter Network
	Spongiform	Not included yet
	Mixed cystic and solid	Interpreter Network
	Solid or almost completely solid	Interpreter Network
ECHO-GENICITY	Anechoic	Measurement by ROI
	Hyperechoic or isoechoic	Measurement by ROI
	Hypoechoic	Measurement by ROI
	Very hypoechoic	Measurement by ROI
SHAPE	Wider-than-tall	Measurement by ROI
	Taller-than-wide	Measurement by ROI
MARGIN	Lobulated or irregular	Interpreter Network
	Smooth	Interpreter Network
	Ill-defined	\
	Extra-thyroidal extension	Not included yet
ECHOGENIC FOCI	None or large comet-tail artifacts	Interpreter Network
	Macrocalcifications	Interpreter Network
	Peripheral calcifications	Not included yet
	Punctate echogenic foci	Interpreter Network

83 matching templates, and the most matching template will be selected as the aided diagnosis report.

C. DIAGNOSTIC AND INTERPRETATION PERFORMANCE OF THE iCADS

To further validate the performance of iCADS, we acquired an experimental data set from 300 participants containing 501 images (208 benign and 293 malignant) as the experiment dataset. We used the results of core needle biopsy (CNB) and fine needle aspiration biopsy (FNAB) as clinical diagnosis for thyroid nodules.

To verify the binary classification performance of the in terms of benign and malignant tumor of thyroid, the corresponding diagnosis result of two experienced radiologist was compared as reference. In addition, another learning based model (VGGNet [29]) which commonly used in assistant diagnosis system is also used for comparison [10], [30]. The precision, recall, accuracy and f1-score of diagnosis were then evaluated. To further validate the correctness of the six interpreters, we compared their results with the judgment of an experienced radiologist (over ten years of experience in ultrasound diagnosis), and calculated related Cohen's kappa coefficient and proportional agreement to analyze the consistency.

D. RADIOLOGIST-COMPUTER COOPERATION EXPERIMENTS

To evaluate the assistance performance of our iCADS, we recruited ten junior radiologists who had 1-3 years of experience at ultrasound diagnosis. They were randomly divided into two equal groups: the experimental group where the five radiologists used the full version of iCADS, which means they were allowed to obtain both the TIRADS matching results and the benignity or malignancy result; and the control group where the radiologists used a short version of iCADS through which they could only obtain the benignity or malignancy result. Since the two groups will receive same binary results, the differences of the final decisions between the two groups would largely depends on the contribution of clinical interpretability that the six interpreters produce.

The comparison was exerted in two steps: 1) The ten junior radiologists independently diagnosed the images in the experimental data set without the help of iCADS; 2) After one week, those junior radiologists were divided into the two groups and used corresponding versions of iCADS to re-diagnose the images in the experimental data set.

We had observed that it takes at least, three seconds to complete a procedure from opening a window and clicking the save button, which means if a diagnosis lasted less than 3 seconds, chances would be high the radiologist had not finish reading iCADS's report. Thus, results that the radiologists were diagnosed less than 3 seconds were automatically deleted. We used precision, recall, accuracy and f1score to compare diagnosis of benign or malignant nodules between the two groups. We employed McNemar's test to compare manual diagnosis with iCAD-aided results in terms of recall and accuracy, and adopted permutation test to compare the changes of precision, recall, accuracy and f1-score in the two groups.

III. RESULT

We conducted a series of experiments to validate the performance of iCAD. We trained the iCAD model using stochastic gradient descent(SGD) method. The training was exerted on a personal computer equipped with NVIDIA GTX 1080Ti. We randomly chose 85% of the image set as the training set and used the remaining 15% of the images as the validation set. The training lasted five hours.

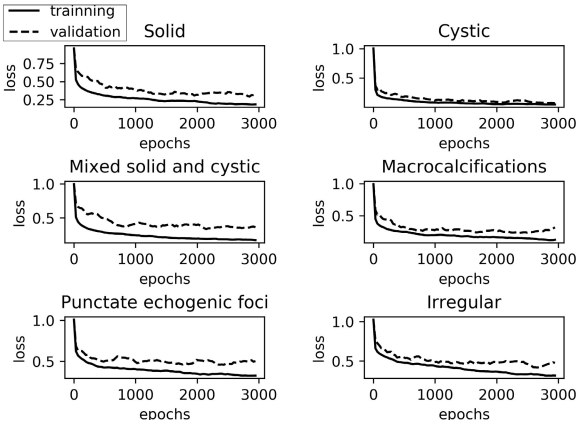


FIGURE 6. The loss curve of each task during the training.

TABLE 2. Inter-observer variability between the CADs and an experienced radiologist.

	Kappa value \Proportional agreement		
	iCADs	S-Detect	SS-FCME
Solid or almost completely solid	0.65\0.86	0.66\0.80	-\0.32
Cystic or almost completely cystic	0.84\0.92		
Mixed cystic and solid	0.57\0.81		
Macrocalcifications	0.63\0.81	0.73\0.87	-\0.48
Punctate echogenic foci	0.61\0.83		
Irregular	0.49\0.75	0.24\0.53	-\0.53

A. PERFORMANCE OF CLINICAL FEATURES EXTRACTION

The effectiveness of the six interpreters is crucial to iCAD's performance. After the training, all the six networks converged to a stable value, as depicted in Fig. 6. As shown in Table. 2, six interpreter networks show substantial or moderate agreement between the CADs and the experienced radiologist in the analysis of inter-observer variability, and their identification results exhibit similar or better performances than the mentioned in previous work [31], [32].

B. PERFORMANCE OF BENIGN AND MALIGNANT CLASSIFICATION

The performances of two networks, including proposed iCADS based on VGG16 [29] and ResNet50 [25] for classification of benign and malignant nodules were compared with experienced radiologists and previous study. As shown in Table. 3, the two networks perform comparably in recall and f1-score, and both are higher than that of the two radiologists, while precisions of the two networks are lower than that of the radiologists. The two neural-network-based iCADS also achieve better results than those of the two methods in previous studies [31], [32].

C. ASSISTANT'S ABILITY

In Fig. 7, no significant difference in accuracy($p > 0.1$), precision($p > 0.1$), recall($p > 0.1$), and f1-score($p > 0.1$) exists between the two groups when exerting independent diagnosis. When the both groups obtained the help of iCADS, whether with TIRADS or not, the recall of both groups were increased (the control group: 71.59%vs. 89.61%, $p = 0.0000$***; the experimental group: 59.51%vs. 92.42%, $p = 0.0000$***), which directly leds to an improvement in accuracy (the control group: 74.05%vs. 77.19%, $p = 0.0555$*; the experimental group: 70.1%vs.80.2%, $p = 0.0000$***) and f1-score (short version icads: 76.36%vs. 82.14%; the experimental group: 69.71%vs. 84.37%). But the precision of the two groups both decreased (the control group: 81.82%vs. 75.82%; the experimental group: 84.14%vs. 77.6%).

TABLE 3. Comparison between different models and experienced radiologists.

	Precision	Recall	Accuracy	F1-score
iCADs(VGG16)	0.81	0.89	0.80	0.85
iCADs(ResNet50)	**0.81**	**0.93**	**0.83**	**0.87**
S-Detect [31]	0.72	0.91	0.81	0.80
SS-FCME [32]	/	/	0.71	/
Experienced Radiologist	0.92	0.69	0.79	0.79

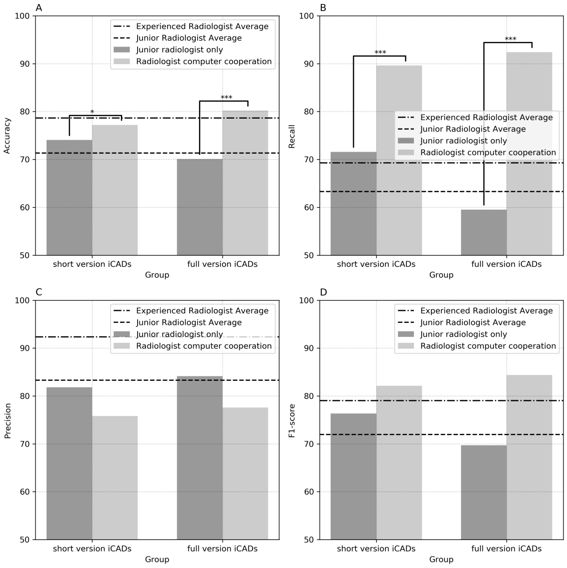


FIGURE 7. Comparison of the diagnostic accuracy(A), recall(B), precision(C),and f1-score(D) between radiologist computer cooperation and junior radiologist only (*p<0.1; **p<0.05;***p<0.01).

The experimental group exhibited significant advantages over the control group: accuracy: 2.24%vs. 9.76%, $p = 0.0159$**; recall: 12.22%vs. 30.99%, $p = 0.0238$**; f1-score: 3.81%vs. 14.35%, $p = 0.0198$**. And the precision of the two groups decreased in a similar level (−4.62%vs. −6.61%, $p = 0.7262$). In particular, the accuracy and the f1-score that the junior radiologists in the experimental group achieved can compete with those of experienced radiologists, which radiologists in the control group could not reach.

IV. DISSCUTION

We designed a neural-network-based approach iCADS that performs better in classification of benign or malignant thyroid nodules and can provide comprehensive assistance to radiologists. We attribute this success to the adaption of

ResNet50 and the adoption of TIRADS as the matching templates. Compared with traditional TIRADS-based classification methods, our iCADS agree better with pathological results; compared with common deep-learning-based methods, iCAD can offer extra useful information for junior radiologists. The experiments demonstrated that the diagnostic evidence presented by iCADS can impose positive influences on radiologists. And our experiments is the first setup that can evaluate the impact of the interpretation in clinical practice.

Most studies on ultrasound-aided diagnosis focused on improving classifiers' performance by training models with large amounts of data or utilizing advanced models, and some physicians acknowledged the value of feature classification, but seldom noticed interpretability of clinical features. We developed an interpretation approach that brings comprehensible clinical features to radiologists for the very first time, and thus exploit potential power of CADs to benefit human-computer cooperation.

Most studies on ultrasound-aided diagnosis focused on improving classifiers' performance by training models with large amounts of data or utilizing advanced models, and some researchers acknowledged the value of feature classification, but seldom noticed interpretability of clinical ultrasound features. Alternatively, we blaze a new trail, clinical features mentioned in the TIRADS were introduced in this study. At present, there are some researches on the clinical features classification. However, previous studies did not report any finding on the radiologist computer cooperation. This study is the first attempt to evaluate the impact of the interpretation in clinical practice.

iCADS's performance on benign and malignant classification indicated that iCADS's precision and recall were higher than those of junior radiologists, which demonstrates iCADS has a potential of providing valuable suggestions for docotors. The performance of clinical feature extraction indicated that the TIRADS matching results generated by the proposed iCADS is consistent with radiologists' understanding of clinical features, which implies that the results of iCADS can be smoothly integrated into the radiologists' reasoning.

The comparison experiments between the experimental group and the control group showed that both groups could increase accuracy,recall,and f1-score but decrease precision from those of the decisions that the junior radiologists made independently. This can account for iCADS itself of holding an ability to provide higher recall and lower precision than radiologists. And this also agrees with common knowledge of AI-based CADs The superiority of iCADS lies in its interpretability. Without six interpreters, iCADS is basically a traditional CAD.

The experiments also indicated that the experimental group provided a higher recall than and a similar precision with those of the control group. The only difference between the two groups was whether they were provided with the six interpreters, which means that the clinical interpretability embedded in iCADS can significantly reduce the rate of misdiagnosis. The control group without TIRADS only received an alert of disagreement or agreement with iCADS, which can remind junior radiologists to make a check on their diagnosis. Basically, compared with the ''disagreement'' alert, it is more important for different radiologists to exchange reasonable ideas from different perspectives in the consultation by human radiologists

CADs naturally hold a different perspective from radiologists, which can benefit doctors' final decision as a ''second opinion''. But its binary output limits the possibility of sharing diagnostic evidence with radiologists. Without clinical interpretation, the CADs's only present benignity and malignancy that is hard to be incorporated into radiologists' reasoning system, especially for junior ones. The simple ''yes or no'' suggestion would also push radiologists to an extreme situation: trusting CADs or ignoring their results.

Our approach that incorporates six interpreters is to establish a bridge between CADs and radiologists. Our experiments on iCADS's assistance ability demonstrated that the bridge is necessary and beneficial to radiologists, by effectively providing detailed interpretation of clinical features.

The interpretation strategy in the proposed iCADs makes it much easier for radiologists to integrate their own opinions into the available evidences, rather than just using either their own results or CADs'. In fact, in the experiment of the radiologists using the short version of iCADS, a large number of samples appeared without effective double reading (operation time less than 3 seconds), which was rare in the experiment of the radiologists using the full version of iCADS.

In order to avoid influences of radiologists' preference or memory, we invited two groups of doctors to use two versions of iCADS, instead of one group used the two versions in turn. Although individual differences might influence the comparison, our experiment on radiologists independent diagnosis showed that no significant differences existed in accuracy, precision, recall, and f1-score between radiologists, implying individual differences can be ignored.

This iCADs were established on a ResNet50 architecture and ACR TIRADS guidelines, but more than just that. Two common deep-learning networks were compared in the classification of benign and malignant nodules: ResNet50 and VGG16, and results showed that they demonstrated acceptable performance, and both showed similar advantages in recall rate, which might allow to complement with radiologists, especially with junior radiologists. This implies our interpretation strategy can be used in other deep-learning-based CADs and obtain similar performance elevation.

Besides, the role of clinical guidelines in this study is to provide the valuable evidences for reasoning. There are different guidelines from the consensus of thyroid diagnosis, which can also be adapted for the interpreter design and training using the proposed strategy.

In conclusion, by introducing the interpretation strategy based on TIRADS guideline, we proposed a novel iCADS with extra clinical feature interpretability for thyroid nodules. A series of experiments demonstrated iCADS can improve junior radiologists' diagnosis decisions remarkably by presenting them comprehensible clinical feature matching. Our study is the first attempt to investigate the effect of interpretability of CADs in clinical practice.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA, A Cancer J. Clin.*, vol. 66, no. 1, pp. 7–30, Jan. 2016.

[2] R. K. Lingam, M. H. Qarib, and N. S. Tolley, "Evaluating thyroid nodules: Predicting and selecting malignant nodules for fine-needle aspiration (FNA) cytology," *Insights Into Imag.*, vol. 4, no. 5, pp. 617–624, May 2013.

[3] J. D. Iannuccilli, J. J. Cronan, and J. M. Monchik, "Risk for malignancy of thyroid nodules as assessed by sonographic criteria," *J. Ultrasound Med.*, vol. 23, no. 11, pp. 1455–1464, Nov. 2004.

[4] M. L. Giger, N. Karssemeijer, and J. A. Schnabel, "Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer," *Annu. Rev. Biomed. Eng.*, vol. 15, no. 1, pp. 327–357, Jul. 2013.

[5] S. Katsigiannis, E. G. Keramidas, and D. Maroulis, "A contourlet transform feature extraction scheme for ultrasound thyroid texture classification," *Int. J. Eng. Intell. Syst. Elect. Eng. Commun.*, vol. 18, nos. 3–4, p. 171, 2012.

[6] M. A. Savelonas, D. K. Iakovidis, N. Dimitropoulos, and D. Maroulis, "Computational characterization of thyroid tissue in the radon domain," in *Proc. 20th IEEE Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2007, pp. 189–192.

[7] Y. Chang, A. K. Paul, N. Kim, J. H. Baek, Y. J. Choi, E. J. Ha, K. D. Lee, H. S. Lee, D. Shin, and N. Kim, "Computer-aided diagnosis for classifying benign versus malignant thyroid nodules based on ultrasound images: A comparison with radiologist-based assessments," *Med. Phys.*, vol. 43, no. 1, pp. 554–567, Jan. 2016.

[8] T. Liu, S. Xie, J. Yu, L. Niu, and W. Sun, "Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 919–923.

[9] S. Han, H.-K. Kang, J.-Y. Jeong, M.-H. Park, W. Kim, W.-C. Bang, and Y.-K. Seong, "A deep learning framework for supporting the classification of breast lesions in ultrasound images," *Phys. Med. Biol.*, vol. 62, no. 19, pp. 7714–7728, Sep. 2017, doi: 10.1088%2F1361-6560%2Faa82ec.

[10] X. Liu, J. Song, S. Wang, J. Zhao, and Y. Chen, "Learning to diagnose cirrhosis with liver capsule guided ultrasound image classification," *Sensors*, vol. 17, no. 12, p. 149, Jan. 2017. [Online]. Available: http://europepmc.org/articles/PMC5298722

[11] Y. Zhu, L. Wang, M. Liu, C. Qian, A. Yousuf, A. Oto, and D. Shen, "MRI-based prostate cancer detection with high-level representation and hierarchical classification," *Med. Phys.*, vol. 44, no. 3, pp. 1028–1039, Mar. 2017.

[12] X. Li *et al.*, "Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: A retrospective, multicohort, diagnostic study," *Lancet Oncol.*, vol. 20, no. 2, pp. 193–201, Feb. 2019.

[13] B. Zhang, J. Tian, S. Pei, Y. Chen, X. He, Y. Dong, L. Zhang, X. Mo, W. Huang, S. Cong, and S. Zhang, "Machine Learning–assisted system for thyroid nodule diagnosis," *Thyroid*, vol. 29, no. 6, pp. 858–867, Jun. 2019.

[14] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Jan. 2017.

[15] J. H. Chen and S. M. Asch, "Machine learning and prediction in medicine–beyond the peak of inflated expectations," *New England J. Med.*, vol. 376, no. 26, pp. 2507–2509, Jun. 2017, doi: 10.1056/NEJMp1702071.

[16] Z. Obermeyer and E. J. Emanuel, "Predicting the future–big data, machine learning, and clinical medicine," *New England J. Med.*, vol. 375, no. 13, pp. 1216–1219, Sep. 2016. [Online]. Available: http://europepmc.org/articles/PMC5070532

[17] N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman, "Deep convolutional networks do not classify based on global object shape," *PLOS Comput. Biol.*, vol. 14, no. 12, Dec. 2018, Art. no. e1006613.

[18] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–10.

[19] M. Schott, "An ultrasonogram reporting system for thyroid nodules stratifying cancer risk for clinical management," *Yearbook Endocrinology*, vol. 2010, pp. 147–149, Jan. 2010.

[20] B. R. Haugen, E. K. Alexander, K. C. Bible, G. M. Doherty, S. J. Mandel, Y. E. Nikiforov, F. Pacini, G. W. Randolph, A. M. Sawka, M. Schlumberger, K. G. Schuff, S. I. Sherman, J. A. Sosa, D. L. Steward, R. M. Tuttle, and L. Wartofsky, "2015 american thyroid association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: The american thyroid association guidelines task force on thyroid nodules and differentiated thyroid cancer," *Thyroid*, vol. 26, no. 1, pp. 1–133, Jan. 2016.

[21] F. N. Tessler, W. D. Middleton, E. G. Grant, J. K. Hoang, L. L. Berland, S. A. Teefey, J. J. Cronan, M. D. Beland, T. S. Desser, M. C. Frates, L. W. Hammers, U. M. Hamper, J. E. Langer, C. C. Reading, L. M. Scoutt, and A. T. Stavros, "ACR thyroid imaging, reporting and data system (TI-RADS): White paper of the ACR TI-RADS committee," *J. Amer. College Radiol.*, vol. 14, no. 5, pp. 587–595, May 2017.

[22] L. Gao, X. Xi, Y. Jiang, X. Yang, Y. Wang, S. Zhu, X. Lai, X. Zhang, R. Zhao, and B. Zhang, "Comparison among TIRADS (ACR TI-RADS and KWAK- TI-RADS) and 2015 ATA guidelines in the diagnostic efficiency of thyroid nodules," *Endocrine*, vol. 64, no. 1, pp. 90–96, Jan. 2019.

[23] Q. Huang, Y. Chen, L. Liu, D. Tao, and X. Li, "On combining biclustering mining and AdaBoost for breast tumor classification," *IEEE Trans. Knowl. Data Eng.*, to be published.

[24] Q.-S. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: A survey," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 27–39, Jan. 2018.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[27] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[28] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2018, pp. 1930–1939.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR* , 2014, pp. 1–14.

[30] K. Yan, M. Bagheri, and R. M. Summers, "3D context enhanced region-based convolutional neural network for end-to-end lesion detection," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2018, pp. 511–519.

[31] Y. J. Choi, J. H. Baek, H. S. Park, W. H. Shim, T. Y. Kim, Y. K. Shong, and J. H. Lee, "A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: Initial clinical assessment," *Thyroid*, vol. 27, no. 4, pp. 546–552, Apr. 2017, doi: 10.1089/thy.2016.0372.

[32] W. Hao, Y. Yang, P. Bo, and C. Qin, "A thyroid nodule classification method based on TI-RADS," *Proc. SPIE Soc. Photo-Opt. Instrum. Eng.*, vol. 10420, Jul. 2017, Art. no. 1042041, doi: 10.1117/12.2281600.

SHIJIE ZHANG received the B.S. degree in engineering from Xidian University, China, in 2015. He is currently pursuing the Ph.D. degree in biomedical engineering with Peking University. He has published two SCI articles in computer-aided diagnosis systems. His research interests are in the area of computer aided diagnosis based on ultrasound imaging.

HUARUI DU received the B.S. degree in mathematics from Shandong University, China, in 2010, and the Ph.D. degree in engineering from Peking University, China, in 2017.

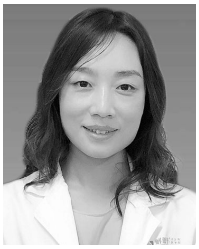
His research interests include medical signals, image analysis, indoor navigation, ultrasound and magnetic field.

ZHUANG JIN received the M.D. degree from the Fourth Military Medical University, China, in 2009. He is currently pursuing the Ph.D. degree in imaging medicine with the Medical School of Chinese PLA, Beijing, China. His research interests are in the areas of Superficial and abdominal ultrasound.

YAQIONG ZHU received the M.D. degree in rehabilitation medicine from the Medical School of Chinese PLA, Beijing, China, in 2016, where she is currently pursuing the Ph.D. degree in imaging medicine.

Her research interest includes the development of artificial intelligence (AI) in ultrasound diagnostics such as identification of benign and malignant thyroid nodules with AI and multimodal ultrasonography diagnosis in skeletal muscle disease.

YING ZHANG received the bachelor's degree in clinical medicine from the Shanxi Medical University, China, in 2015, and the master's degree in imaging and nuclear medicine from the PLA Medical College, China, in 2018. She is currently pursuing the Ph.D. degree in imaging and nuclear medicine with the Medical School of Chinese PLA, Beijing, China. She has published one article in ultrasonic diagnosis. Her research interest is in the area of ultrasonic diagnosis of thyroid diseases.



FANG XIE received the Ph.D. degree in rehabilitation medicine from the Medical School of Chinese PLA, Beijing. She is currently an Associated Chief Physician with the Department of Ultrasound, General Hospital of Chinese PLA. She is currently an Associate Chief Physician and Ultrasound Specialist of the China Medical Imaging Quality Research Committee. She is the author of one book and four articles. Her research interest is in the area of artificial intelligence in ultrasound diagnostics.

MINGBO ZHANG received the M.D. degree from Health Science Center, Peking University, in 2010. She is currently an Associate Chief Physician with the Department of Ultrasound, General Hospital of the People's Liberation Army. She has rich experience in the ultrasound diagnosis of thyroid nodules. She is also an expert in ultrasound guided biopsy and ablation of thyroid nodules. She has published three SCI articles in the thyroid field and gave a speech in the America Thyroid Association (ATA) Conference, in 2017. She received the Outstanding Chinese Young Ultrasound Doctor Prize, in 2016, the First Prize in the First National AI-doctor Diagnosis Competition of Thyroid Nodules, in 2017, and the New Talent Prize of PLA General Hospital, in 2017.

XIAOQI TIAN received the M.D. degree in pathology from Tianjin Medical University Tianjin, China, in 2014. She is currently pursuing the Ph.D. degree in imaging medicine with the Medical School of Chinese PLA, Beijing, China. Her research interest includes the evaluation of the efficacy of thyroid cancer ablation and multimodal ultrasonography diagnosis in kidney.

JUE ZHANG received the Ph.D. degree in engineering mechanics from Peking University, Beijing, China, in 2003. Since 2009, he has been an Assistant Professor with the Academy for Advanced Interdisciplinary Studies, Peking University. He is the author of more than 200 articles. His research interests include medical signals, image analysis, clinical fMRI pulse sequence, RF coil design, plasma medicine, and nanosecond pulsed electric field.

YUKUN LUO received the Ph.D. degree in medical imaging from the Medical School of Chinese PLA, Beijing, China, in 2005. Since 2011, she has been a Chief Physician with the Department of Ultrasound, General Hospital of the People's Liberation Army. She is the author of more than 60 articles. Her research interests include cervical metastatic lymph nodes from papillary thyroid carcinoma, contrast-enhanced ultrasound enhancement patterns for thyroid nodules, and ultrasound-guided radiofrequency ablation for treating.