International Journal of Knowledge and Systems Science
Volume 13 • Issue 1

Diagnostic Performance of Artificial Intelligence for Interpreting Thyroid Cancer in Ultrasound images

Piyanuch Arunrukthavon, Mahidol University, Thailand
Dittapong Songsaeng, Mahidol University, Thailand
Chadaporn Keatmanee, Ramkhamhaeng University, Thailand*
Songphon Klabwong, Asian Institute of Technology, Thailand
Mongkol Ekpanyapong, Asian Institute of Technology, Thailand
Matthew N. Dailey, Asian Institute of Technology, Thailand

ABSTRACT

Thyroid ultrasonography is mainly used for the detection and characterization of thyroid nodules. However, there is some limitation since the diagnostic performance remains highly subjective and depends on radiologist experiences. Therefore, artificial intelligence (AI) was expected to improve the diagnostic performance of thyroid ultrasound. To evaluate the diagnostic performance of the AI for differentiating malignant and benign thyroid nodules and compare it with that of an experienced radiologist and a third-year diagnostic radiology resident, 648 patients with 650 thyroid nodules, who underwent thyroid ultrasound guided-FNA biopsy and had a decisive diagnosis from FNA cytology at Siriraj Hospital between January 2014 and June 2020, were enrolled. Although the specificity and accuracy were slightly higher in AI than the experienced radiologist and the resident (specificity 78.85% vs. 67.31% vs. 69.23%; accuracy 78.46% vs. 70.77% vs. 70.77%, respectively), the AI showed comparable diagnostic sensitivity and specificity to the experienced radiologist and the resident (p=0.187-0.855).

KEYWORDS

Convolutional Neural Network, Deep Learning, Machine Learning, Medical Image Processing, Thyroid Cancer, Thyroid Nodule Classification, Ultrasound Images

INTRODUCTION

Thyroid nodules are common in the general population with a prevalence of 20-60% (Dean & Gharib, 2008) and can be either malignant or benign. The etiologies of thyroid nodules are a simple overgrowth of normal thyroid tissue, inflammation, or tumor. Thyroid cancer is one of the most common types of cancer in the endocrine system. It is the fifth most common cancer of women worldwide and the fourth most common cancer of women in Thailand (Pellegritit et al., 2013). Recent research showed

DOI: 10.4018/IJKSS.309431 *Corresponding Author

that the annual incidence of thyroid cancer has gradually increased and now accounts for 1,500 new patients per year in Thailand (Tangjaturonrasme et al., 2018). Papillary thyroid carcinoma, the predominant subtype of thyroid cancer showed an increase in incidence in Thailand in a decade. (Bychkov, 2017; Bychkov, et al., 2017).

Nowadays, there are many diagnostic imaging modalities to detect thyroid nodules, including ultrasonography, computed tomography, magnetic resonance imaging, positron emission tomography, and scintigraphy. In clinical practice, thyroid ultrasonography is mainly used for both the detection and characterization of thyroid nodules. It is preferable due to noninvasiveness, convenience, no radiation exposure, and relatively low price, and intervention procedures can be performed concurrently. However, ultrasonography has some limitations to differentiate benign thyroid nodules from malignant ones due to the complex structures of thyroid nodules. Consequently, thyroid ultrasonography remains highly subjective and depends on physicians' experience, which causes a greater risk of misdiagnosing cancer and increases the number of FNA biopsies.

In the healthcare field, novel technologies have been developed in many countries for supporting a diagnosis such as medical records, medical text analysis, interprofessional team collaboration, and AI. These electronic medical records (EMRs) are useful for data sharing among medical departments (Taewijit & Theeramunkong, 2021). Medical text analysis uses knowledge management for disease prediction (Menaouer et al., 2020). The systems approach Interprofessional Team Collaboration (IPC) enhances outcomes of healthcare services, as well as improves the safety and quality of healthcare setups (Matsushita, et al. 2021). AI is used for diagnostic and therapeutic purposes in medical imaging. AI has shown impressive accuracy and sensitivity in the identification of imaging abnormality and tissue characterization. Thus, AI is expected to play an essential role in assisting radiologists in characterizing thyroid nodules. This can reduce errors caused by subjective factors, assist the diagnostic performance in avoiding unnecessary FNA biopsies, benefit further treatment plans for patients, and reduce healthcare costs.

Computer-aided diagnosis (CAD) has been considered an innovation in modern healthcare (Keung et al., 2018; Matsushita et al., 2018; Rathi & Pareek, 2019). CAD can effectively support in different domains including Thyroid cancer screening. There are three main parts of CAD, (a) thyroid cancer boundary segmentation, (b) significant feature analysis, and (c) classification. In this work, the thyroid nodule classification part is the focus. Basically, thyroid nodules have variations in size, shape, echogenicity, composition, and calcification. Therefore, the simple extracted features applied in the conventional CAD algorithms could limit distinction because of the simplicity. The different approaches, deep learning models, for example, the convolutional neural networks (CNNs) significantly were applied in knowledge management in identifying the appropriate prediction (Kengpol & Punyota, 2022). Moreover, in the past decades, deep CNNs have shown an impressive outcome in various machine vision contributions including in medical image processing such as object segmentation and feature classification (Khachnaoui & Khlifa, 2018).

MATERIALS

Patients

648 patients with 650 thyroid nodules, who underwent thyroid ultrasonography with ultrasound-guided FNA biopsy at Siriraj Hospital between January 2014 and June 2020, were enrolled in this retrospective study. The inclusion criteria were as follows: (a) age equal to or more than 18 years old, (b) underwent ultrasound thyroid with ultrasound-guided FNA biopsy within the same day, and (3) had a decisive diagnosis of benign or malignant thyroid nodule by thyroid FNA cytology. According to the authors' radiology database, thyroid nodules with benign cytology account for 80-90% of overall thyroid nodules .Therefore, we randomly selected the benign case in proportion with the malignant

case for each year. Patients were excluded from the study if (a) the thyroid nodules had no clear diagnostic findings from pathology or (b) had suboptimal ultrasound image quality.

Ultrasound Image Acquisition and Analysis

In this study, thyroid ultrasound examinations were performed using a high-frequency linear probe by experienced radiologists in the thyroid imaging center. The thyroid gland was scanned in both transverse and longitudinal planes with grayscale and color Doppler techniques based on the American College of Radiology accreditation standard. Then the images were stored on a picture archiving and communication system (PACS). There were 650 thyroid nodules in total, among which 130 were malignant, and 520 were benign thyroid nodules. The majority of images included were randomly assigned to the training, validation, and test groups, 80%, 10%, and 10% respectively.

The authors collected the ultrasound images of thyroid nodules as grayscale images on the picture archiving and communication system (PACS) and exported them in a PowerPoint program. The 10 years' experienced radiologist and the third-year diagnostic radiology resident evaluated the ultrasound thyroid images for each nodule included the following: size (maximum diameter in cm), echogenicity (marked hypoechoic, hypoechoic, isoechoic, hyperechoic), composition (predominant cyst, predominant solid, solid), calcification (microcalcification, macrocalcification, none), shape (taller, wider) and margin (well-defined, ill-defined). The ACR Thyroid Imaging Reporting and the Data System (ACR TI-RADS) guidelines were used to differentiate each thyroid nodule based on its ultrasound features as benign or malignant.

METHODOLOGY

A thyroid nodule on a grayscale ultrasound image modality was collected when it was viewed using a PowerPoint program. The dataset was retrospectively selected for each representative image, as shown in Figure 1. After that, the thyroid nodule region was systematically cropped on the representative image of each nodule. The ultrasound images were statistical randomly assigned to the training, validation, and test groups, 80%, 10%, and 10% respectively. The overall workflow was shown in Figure 2.

Figure 2. shows the flow pre-processing techniques for ultrasound images before putting them into the AI model. Because of the number of input images, the ROI should be segmented from the original ultrasound images shown in Figure 1. The segmented region relied on the marks indicating the boundary of a thyroid nodule which were located by radiologists. The cropping process was done manually. After that, the image inpainting was performed to remove the marks. Finally, the ultrasound images were resized. In the example, the input image was resized as 180x180 pixels. Whereas the size of the input images was specified by the experiment with the potential candidate AI models which is explained in the section: Fine-Tuning the Hyper-Parameter of AI Models.

Figure 1. (A) Malignant thyroid nodule and (B) Benign thyroid nodule

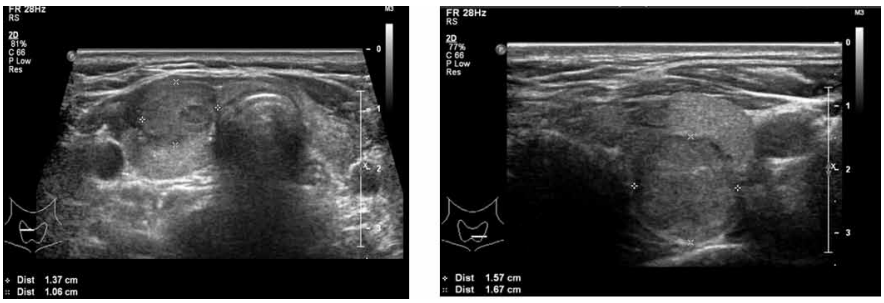
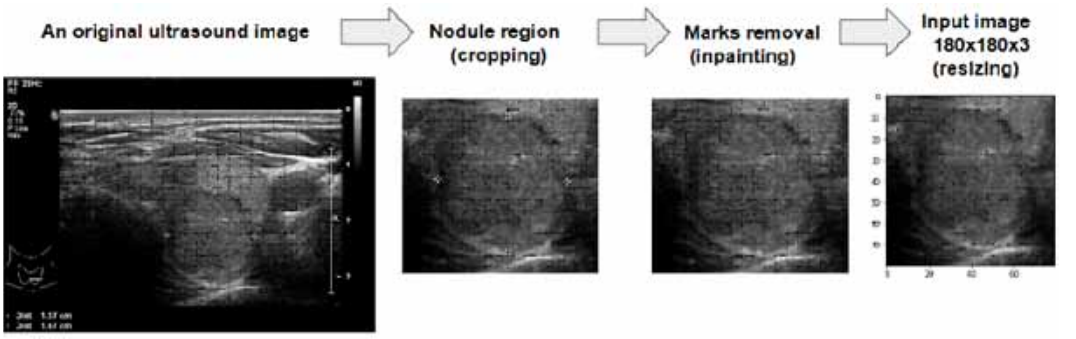


Figure 2. Pre-processing techniques for input images



The workflow for the AI model used in this paper as shown in Figure 3. depicts the end-to-end process for predicting the types of thyroid nodules. The pre-processed ultrasound image is the input. Despite the small data set, image augmentation was used to increase the training group's sample size. After that, a pre-trained model was used to predict the thyroid nodule in the input image as the output. The prediction result was either benign or malignant.

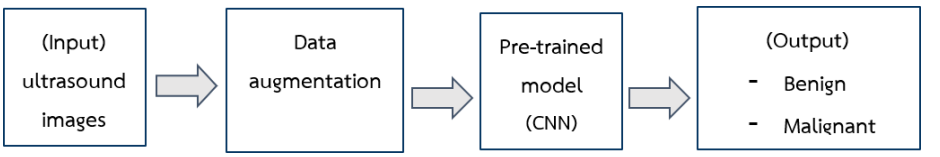
In this study, the transfer learning method was employed due to the limitation of dataset size. One common approach for transfer learning is to use the pre-trained model (Torrey & Shavlik, 2010), in which the model is pre-trained with data from the different domains. In this case, the authors used the pre-trained model based on the ImageNet dataset and subsequently trained with datasets from the hospital. There were four pre-trained models as potential candidates; VGG16 (Simonyan & Zisserman, 2014), InceptionV3 (Szegedy et al., 2016), DenseNet121 (Huang et al., 2017), and EfficientNet (Tan & Le, 2019) were transferred with the binary classification to distinguish between benign and malignant thyroid nodules on the top of their structure.

VGG16 is a classical CNN model. The network uses small 3x3 kernels and pooling sized 2x2 for the convolutional layers which were consistently arranged throughout the whole architecture. The other components are two fully connected layers followed by a softmax layer. The 16 in VGG16 was referred to the layers that have weights and the total number of parameters is approximately about 138 million parameters.

InceptionV3 is a fewer parameters CNN model composed of 42 layers. The upgraded inception model version 3 was done by reducing the dimensions to avoid the drastic alternation of the input dimensions for better performance, utilizing smart factorization methods for enhancing computational efficiency, regularizing by having Batchnorm or Dropout operations for increasing contribution of the auxiliary classifiers.

DenseNet121 is a Deeper CNN using the architectures such as highway networks, residual networks, and fractal networks in the direct connectivity pattern. Its design aimed to maximize information and gradient flow from the input layer until the output layer. Moreover, to directly connect every layer makes the network not need to learn redundant feature maps, thus the parameters can be reduced.

Figure 3. The overall workflow of AI



EfficientNet is a scaling-up CNN model. It uses a technique named compound coefficient to increase its accuracy and efficiency. The compound scaling method is the balancing dimensions in width, depth, and resolution by scaling with a constant ratio. The model was developed using 7 models with various dimensions for compound scaling.

Fine-Tuning the Hyper-Parameter of AI Models

These four state-of-the-art models were evaluated with the training and validation groups. The basic hyper-parameters of the CNN pre-trained models were adjusted for fine-tuning the models. There are a lot of numbers in the hyper-parameter adjustment, but the authors had chosen only the values that demonstrated significant results.

The first hyper-parameter was the size of the input images that control the fixed size of the input layer for the pre-trained models. Although the pre-processed images were segmented in different sizes based on the boundary of thyroid nodules, the shape of the input images for the pre-trained models had to be fixed. In this study, the shape of the input images was two-dimensional colored images, thus, the size of the input would be *image width* x *image height* x *channel(color)*. In this experiment, the fine adjustment of the input size parameter was examined, however, there were three significantly different results shown in Table 1. From the table, the pre-trained model which has the highest performance was DenseNet121 with *180 x 180 x 3* for the size of the input image parameter.

The second hyper-parameter was the batch size that controlled the number of training samples to work through before the pre-trained models' internal parameters were updated. As the size of the input parameter, there were three significantly different results, shown in Table 2. From the table, the pre-trained model with the highest performance was DenseNet121 with the batch size being 8.

The last hyper-parameter was the number of epochs that controlled the number of complete passes through the training group. Table 3 shows three significantly different results. From the table, the pre-trained model with the highest performance was DenseNet121 with 200 epochs.

Table 1. The evaluation of the pre-train models by adjusting the size of the input images parameter

Pre-Train Model	Size of the Input Image (*Pixel* x *Pixel* x *Channel*)					
	280 x 280 x 3		180 x 180 x 3		80 x 80 x 3	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
VGG16	74.78	74.07	75.23	74.38	73.78	73.38
InceptionV3	73.75	73.45	73.22	73.12	72.89	72.01
DenseNet121	78.87	77.76	81.10	80.21	75.65	75.31
EfficientNet (B0)	69.29	65.19	68.06	65.17	68.72	65.11

Table 2. The evaluation of the pre-train models by adjusting the batch size

Pre-Train Model	Batch Size (# of Training Samples)					
	8		32		128	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
VGG16	75.23	74.38	63.08	62.87	60.55	60.01
InceptionV3	73.22	73.12	69.55	67.98	60.68	59.99
DenseNet121	81.10	80.21	77.15	75.13	71.09	70.11
EfficientNet (B0)	68.06	65.17	65.98	64.11	60.45	60.00

Table 3. The evaluation of the pre-train models by adjusting the number of epochs

Pre-Train Model	The Number of Epochs					
	50		200		800	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
VGG16	70.11	69.05	75.23	74.38	73.65	72.87
InceptionV3	70.34	70.00	73.22	73.12	71.39	70.65
DenseNet121	75.23	14.22	81.10	80.21	78.87	77.08
EfficientNet (B0)	65.88	63.89	68.06	65.17	66.88	65.76

After considering the experimental results from Table 1 to Table 3, the pre-trained model named DenseNet121 overcame others. Therefore, the DenseNet12 was utilized for further experiments.

Data Augmentation

According to the few numbers of the training group, the data augmentation was randomly conducted according to batch size setting. The examples of data augmentation are displayed in Figure 4. There were three techniques used, rotation, zoom, and shear. The range of rotation, zoom, and shear was randomly set between 0 to 20 degrees, 0 to 20 percent, and 0 to 10 respectively. The effectiveness of adding augmented data is shown in Table 4. The pre-trained model with data augmentation had better performance compared to the none.

From previous experimental results, the pretrain model, DenseNet121 was applied to help diagnose thyroid nodules as benign or malignant thyroid nodules. After that, prediction of AI was compared to the blinded experienced diagnostic radiologist and the third-year diagnostic radiology resident based on US images on the test group.

Figure 4. Example of data augmentation

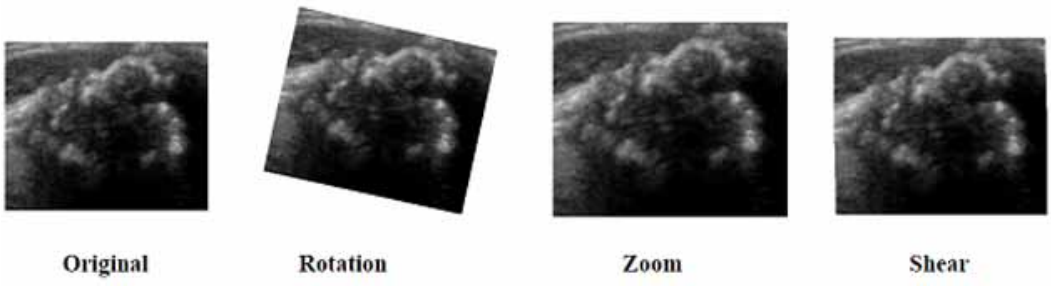


Table 4. The evaluation of the data augmentation

Data augmentation	Rotation, Zoom, & Shear	
	Sensitivity	Specificity
No	73.01	72.22
Yes	81.10	80.21

Outcome Measurement

The primary outcome was the diagnostic performance of AI for the diagnosis of malignant and benign thyroid nodules. The secondary outcome was the diagnostic performance of the AI compared with that of the experienced radiologist and third-year diagnostic radiology resident.

Data and Statistical Analysis

All statistical analyses were done by using SPSS software version 26. The continuous data were presented in means, standard deviation, and number, while the categorical data were presented in percentages. The diagnostic performance of thyroid nodules was presented as sensitivity, specificity, accuracy, positive predictive value, and negative predictive value. For categorical data between benign and malignant thyroid nodules, the Chi-square test and independent T-test were used to compare the two. The Cohen's Kappa statistic was used to compare the interobserver agreement of the final diagnosis of AI, the experienced radiologist, and the resident.

RESULTS

Demographic Data

There were 650 thyroid nodules from 648 patients with a mean age of 54.89±14.19 years (range 18-86 years). Female patients (n=568, 87.7%) were predominant in the study (Table 5). Of the 650 nodules, 520 (80%) were found by pathology to be benign, and 130 (20%) were found to be malignant (Table 5). The mean size of nodules was 1.8 cm (range 0.1-6.0 cm), and most of the patients had a single nodule (30.2%) (Table 6).

In the test group, there were 65 nodules from 65 patients with a mean age of 51.42±16.18 years (range 18-83 years). Of 65 nodules, 52 (80%) were found by pathology to be benign, and 13 (20%) were found to be malignant. The patient's age and nodule size were significantly different between benign

Table 5. Demographic data of all patients and the test data set

Variables	Number (%) or Mean ±SD	
	All Data Set 650 Nodules	The Test Set 65 Nodules
Age (years)	54.89±14.19	51.42±16.18
Male	80 (12.3)	10 (15.4)
Female	568 (87.7)	55 (84.6)
Size (cm)	1.8 (0.1-6.0)	2.3 (0.4-6.0)
Number of Nodules		
1	196 (30.2)	21 (32.3)
2	140 (21.6)	12 (18.5)
3	90 (13.9)	10 (15.4)
4	70 (10.8)	7 (10.8)
5	51 (7.9)	3 (4.6)
>5	101 (15.6)	12 (18.5)
Benign Nodules	520 (80)	52 (80)
Malignant Nodules	130 (20)	13 (20)

Table 6. Comparison of demographic data of benign and malignant thyroid nodules in the test group

Variables	Total	Benign (52)	Malignant (13)	P-value
Age	51.42±16.18	55.17±15.23	36.38±10.23	P<0.001
Male **Female**	10 (15.4%) 55 (84.6%)	7 (13.5%) 45 (86.5%)	3 (23.1%) 10 (76.9%)	0.405
Nodule Size	2.3 (0.4-6.0)	2.4 (0.5-6.0)	0.9 (0.4-2.0)	P<0.001
No. of Nodule **1** **2** **3** **4** **5** **>5**	 21 (32.3%) 12 (18.5%) 10 (15.4%) 7 (10.8%) 3 (4.6%) 12 (18.5%)	 19 (36.5%) 11 (21.2%) 7 (13.5%) 3 (5.8%) 2 (3.8%) 10 (19.2%)	 2 (15.4%) 1 (7.7%) 3 (23.1%) 4 (30.8%) 1 (7.7%) 2 (15.4%)	0.078

and malignant groups. (p<0.001, p<0.001, respectively) The demographic data such as gender and number of nodules showed no statistical difference between benign and malignant groups (Table 6).

Nodule Characteristics

The ultrasound features of overall thyroid nodules in the test group were as shown in Table 7. The majority of the ultrasound characteristics which were significantly associated with malignant thyroid nodules were microcalcification (p<0.001) and a taller shape (p=0.003). In contrast, benign nodules tended to present with wider shapes and no calcification. Most benign and malignant groups also had solid compositions. No significant difference in the echogenicity and margin was seen comparing benign and malignant groups.

Table 7. Comparison of ultrasound features of benign and malignant thyroid nodules in the test group

Variables	Benign Nodules (n=52)	Malignant Nodules (n=13)	P-Value
Echogenicity **Marked Hypoechoic** **Hypoechoic** **Isoechoic** **Hyperechoic**	 10 (19.2%) 12 (23.1%) 18 (34.6%) 12 (23.1%)	 4 (30.8%) 6 (46.2%) 1 (7.7%) 2 (15.4%)	0.127
Composition **Solid** **Predominant Solid** **Predominant Cyst**	 28 (53.8%) 16 (30.8%) 8 (15.4%)	 12 (92.3%) 1 (7.7%) -	0.047
Calcification **Microcalcification** **Macrocalcification** **None**	 14 (26.9%) 5 (9.6%) 33 (63.5%)	 12 (92.3%) - 1 (7.7%)	P<0.001
Shape **Wider** **Taller**	 43 (82.7%) 9 (17.3%)	 5 (38.5%) 8 (61.5%)	0.003
Margin **Well-defined** **Ill-Defined**	 41 (78.8%) 11 (21.2%)	 11 (84.6%) 2 (15.4%)	1.000

Diagnostic Performance of AI for Differentiating Benign and Malignant Thyroid Nodules

The diagnostic performance values of AI were calculated as shown in Table 8. The sensitivity, specificity, accuracy, positive predictive value, and negative predictive value of AI were 79.92% [46.19-94.96], 78.85% [65.30-88.94], 78.46% [66.51-87.69], 47.62% [33.21-62.43], and 93.18% [83.38-97.38], respectively.

Diagnostic Performance of AI, an Experienced Radiologist, and a Resident for Nodule Classification

The overall diagnostic performance values among the three groups were calculated as shown in Table 9. There was no statistical difference in diagnostic sensitivity and specificity among the AI, an experienced radiologist, and a resident (p=0.187-0.855). The AI showed higher diagnostic specificity and accuracy than an experienced radiologist and a resident without statistical significance (specificity 78.85%, 67.31%, 69.23% and accuracy 78.46%, 70.77%, 70.77%, respectively). The sensitivity of AI was lower than the experienced radiologist and higher than the resident without statistical significance (79.92%, 84.62% and 76.92%, respectively).

Inter-Observer Variability

The summary of inter-observer variability among the AI, the experienced radiologist, and the resident is demonstrated in Table 10. The Kappa value between AI and the experienced radiologist was 0.450 [0.225–0.675], which was interpreted as moderate agreement. The Kappa value between AI and resident was 0.239 [0.000-0.489], which fell under the fair agreement category between the two groups. In contrast, the Kappa value between radiologist and resident was 0.684 [0.503-0.864], showing substantial agreement.

Table 8. Diagnostic performance of AI, the experienced radiologist, and the resident for differentiating benign and malignant thyroid nodules

Diagnostic Values	AI	Experienced Radiologist	Resident
Sensitivity [95%CI]	79.92 [46.19-94.96]	84.62 [54.55-98.08]	76.92 [46.19-94.96]
Specificity [95%CI]	78.85 [65.30-88.94]	67.31 [52.89-79.67]	69.23 [54.90-81.28]
Accuracy [95%CI]	78.46 [66.51-87.69]	70.77 [58.17-81.40]	70.77 [58.17-81.40]
PPV [95%CI]	47.62 [33.21-62.43]	39.29 [29.13-50.46]	38.46 [27.39-50.87]
NPV [95%CI]	93.18 [83.38-97.38]	94.59 [82.83-98.45]	92.31 [81.40-97.05]

Table 9. Comparison of AI, an experienced radiologist, and the resident in terms of diagnostic sensitivity and specificity

	Sensitivity (p-value)	Specificity (p-value)
AI vs. Experienced Radiologist	P=0.725	P=0.187
AI vs. Resident	P=0.855	P=0.266
Experienced Radiologist vs. Resident	P=0.625	P=0.834

Table 10. Inter-observer variability among AI, the radiologist, and the resident for differentiating benign and malignant thyroid nodules

Reader 1	Reader 2	Kappa (95% CI)
AI	Experienced Radiologist	0.450 [0.225–0.675]
AI	Resident	0.239 [0.000-0.489]
Experienced Radiologist	Resident	0.684 [0.503-0.864]

DISCUSSION

Recently, the computer-aided diagnosis system for the diagnosis of thyroid nodules has been studied worldwide. Xu et al. (2020) reported a meta-analysis of 19 studies with 4,781 nodules, which suggested that overall classic machine learning and deep learning-based system demonstrated comparable diagnostic sensitivity and specificity to experienced radiologists (sensitivity 87% [78–93] vs 87% [85–89], specificity 85% [76–91] vs 87% [81–91]). A previous study by Gao et al. (2018) reported that the computer-aided diagnosis system had comparable sensitivity but lower specificity than the experienced radiologists (sensitivity 96.7% vs 96.2% p=0.19) (specificity 48.5% vs. 75.7%, p<0.01). Similarly, a study by Choi et al. (2017) showed the same results (sensitivity 90.7% vs 88.4%, p>0.99) (specificity 74.6% vs 94.9%, p=0.002).

In this study, the authors developed AI from the CNN model to differentiate malignant and benign thyroid nodules. The sensitivity, specificity, and accuracy of the authors' system were 79.92% [46.19-94.96], 78.85% [65.30-88.94], and 78.46% [66.51-87.69], respectively, which were comparable with the previous studies (84.6%, 80%, 88.1%; 80.2%, 82.6%, 81.7%) (Kim et al., 2019; Yoo et al., 2018). In Tables 8 and 9 the AI showed no statistical difference in diagnostic sensitivity and specificity when compared with the experienced radiologist and the resident (p=0.187-0.855). Although, the specificity and accuracy were slightly higher in AI than the experienced radiologist and the resident (Specificity 78.85% [65.30-88.94] vs 67.31 [52.89-79.67] vs 69.23 [54.90-81.28], accuracy 78.46 [66.51-87.69] vs 70.77 [58.17-81.40] vs 70.77 [58.17-81.40], respectively). According to the result, AI might provide added diagnostic value in detecting malignant nodules for the experienced radiologist in clinical practice. This system also assists inexperienced radiologists in making a final diagnosis and avoiding unnecessary FNA biopsies. Although, the reported specificity of the experienced radiologist was relatively low when compared with previous studies (86.4%-95.5%) (Choi et al., 2017; Yoo et al., 2018), the population bias might explain this. All patients in this study underwent FNA biopsies, indicating that the included nodules might have some gray zone features which were difficult to diagnose by ultrasound. The diagnostic sensitivity of AI also showed no statistical significance between the experienced radiologist and the resident. However, we found that AI had higher sensitivity than the resident and lower than the experienced radiologist (79.92%, 84.62% and 76.92%, respectively). Therefore, this system would be useful for ruling out malignant nodules for inexperienced radiologists.

Inter-observer variability in differentiating malignant and benign thyroid nodules showed substantial agreement between the experienced radiologist and the resident (K=0.684). Similarly, all diagnostic performance variables such as sensitivity, specificity, and accuracy for the experienced radiologist and the resident showed no significant difference. In contrast, the AI showed moderate agreement and fair agreement between the experienced radiologist and the resident, respectively (K=0.239, 0.450). This was probably due to the learning process, which had not been elucidated yet.

There were several limitations in our study. First, our sample size was small for the training, validation, and test groups, which influenced the diagnostic performance values. Further studies where the quantity of data is increased by collecting more ultrasound images from different centers and different ultrasound manufacturers would improve this diagnostic performance system. Second, our study was a retrospective study and collected data from a single center in Siriraj hospital. There

could be some selection bias. Third, we included only nodules that had a definitive diagnosis. The undetermined or nondiagnostic cytology was excluded, which would limit the generalizability of the results. Fourth, the radiologist's diagnostic performance was limited by the static images instead of dynamic images, which might have led to misinterpretation according to the Thyroid imaging reporting and data system (TI-RADS) guideline.

CONCLUSION

The AI shows comparable diagnostic performance to the experienced radiologist and the resident. This system also demonstrates good specificity and accuracy, which may have the potential to assist radiologists in diagnosing thyroid cancer. However, there were some effective pre-trained models that have not been investigated such as WSDAN (Weakly Supervised Data Augmentation Network). It is a fine-grained visual classification model which is suitable for distinguishing the thyroid nodules in TI-RADs. Another way to improve the pre-trained model is by obtaining more data from hospitals in Thailand. In addition, to improve healthcare systems in Thailand, developing a thyroid nodule assessment mobile application combined with deep learning together should steer the users (radiologists). They will detect malignancy of the thyroid nodule in no time (Rathi & Pareek, 2019).

ACKNOWLEDGMENT

This work was partially supported by the Department of Computer Science, Faculty of Science, Ramkhamhaeng University, the Department of Radiology, Faculty of Medicine Siriraj Hospital, Mahidol University, and the AI Center, Asian Institute of Technology, Thailand. According to protocol approval from our institutional review board in ID SIRB Protocol No. 851/2563(IRB4), informed consent is waived for the retrospective study. Anonymization was applied to the patient's information and ultrasound images.

CONFLICT OF INTEREST

The authors of this publication declare there is no conflict of interest.

FUNDING AGENCY

This research was partially supported by the Broadcasting and Telecommunications Research and Development Fund for Public Interest [grant number A63-1-(2)-018].

REFERENCES

Bychkov, A., Hirokawa, M., Jung, C. K., Liu, Z., Zhu, Y., Hong, S. W., Saton, S., Lai, C., Huynh, L., & Kakudo, K. (2017). Low rate of noninvasive follicular thyroid neoplasm with papillary-like nuclear features in Asian practice. *Thyroid*, *27*(7), 983–984. doi:10.1089/thy.2017.0079 PMID:28486057

Bychkov, A. A. (2017). Pathologist's perspective on thyroid cancer trends in Thailand. *Cancer Epidemiology*, *47*, 133–134. doi:10.1016/j.canep.2017.02.009 PMID:28274695

Choi, Y. J., Baek, J. H., Park, H. S., Shim, W. H., Kim, T. Y., Shong, Y. K., & Lee, J. H. (2017). A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: Initial clinical assessment. *Thyroid*, *27*(4), 546–552. doi:10.1089/thy.2016.0372 PMID:28071987

Dean, D. S., & Gharib, H. (2008). Epidemiology of thyroid nodules. *Best Practice & Research. Clinical Endocrinology & Metabolism*, *22*(6), 901–911. doi:10.1016/j.beem.2008.09.019 PMID:19041821

Gao, L., Liu, R., Jiang, Y., Song, W., Wang, Y., Liu, J., Wang, J., Wu, D., Li, S. H. A., & Zhang, B. (2018). Computer-aided system for diagnosing thyroid nodules on ultrasound: A comparison with radiologist-based clinical assessments. *Head & Neck*, *40*(4), 778–783. doi:10.1002/hed.25049 PMID:29286180

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2017*, IEEE, 4700-4708.

Kengpol, A., & Punyota, W. (2022). Knowledge management of vegetarian food for the elderly using DCNN: An empirical study in Thailand. *International Journal of Knowledge and Systems Science (IJKSS)*, 13(2), 1-17.

Keung, K. L., Lee, C., Ng, K. K. H., Leung, S. S., & Choy, K. L. (2018). An empirical study on patients' acceptance and resistance towards electronic health record sharing system: A case study of Hong Kong. *International Journal of Knowledge and Systems Science (IJKSS)*, *9*(2), 1-27.

Khachnaoui, H., Guetari, R., & Khlifa, N. (2018). A review on deep learning in thyroid ultrasound computer-assisted diagnosis systems. *IEEE International Conference on Image Processing, Applications and Systems (IPAS)*, 2018, IEEE, 291-297.

Kim, H. L., Ha, E. J., & Han, M. (2019). Real-world performance of computer-aided diagnosis system for thyroid nodules using ultrasonography. *Ultrasound in Medicine & Biology*, *45*(10), 2672–2678. doi:10.1016/j.ultrasmedbio.2019.05.032 PMID:31262524

Matsushita, H., Lillrank, P., & Ichikawa, K. (2018). Human competency as a catalyzer of innovation within health and nursing care through a perspective of complex adaptive systems. *International Journal of Knowledge and Systems Science*, *9*(4), 1–15. doi:10.4018/IJKSS.2018100101

Matsushita, H., Orchard, C., Fujitani, K., & Ichikawa, K. (2021). A systems study on interprofessional collaboration in healthcare: Testing the Japanese version of the Assessment of Interprofessional Team Collaboration Scale II. *International Journal of Knowledge and Systems Science*, *12*(3), 93–108. doi:10.4018/IJKSS.2021070105

Menaouer, B., Mohammed, S., & Nada, M. (2020). Towards a model to improve Boolean knowledge mapping by using text mining and its applications: Case study in healthcare. *International Journal of Information Retrieval Research*, *10*(3), 35–56. doi:10.4018/IJIRR.2020070103

Pellegriti, G., Frasca, F., Regalbuto, C., Squatrito, S., & Vigneri, R. (2013). Worldwide increasing incidence of thyroid cancer: Update on epidemiology and risk factors. *Journal of Cancer Epidemiology*, *1*, 965212. PMID:23737785

Rathi, M., & Pareek, V. (2019). Mobile based healthcare tool an integrated disease prediction & recommendation system. *International Journal of Knowledge and Systems Science*, *10*(1), 38–62. doi:10.4018/IJKSS.2019010103

Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. arxiv.org.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818-2826. doi:10.1109/CVPR.2016.308

Taewijit, S., & Theeramunkong, T. (2021). Learning pattern relation-based hyperbolic embedding for adverse drug reaction extraction. *International Journal of Knowledge and Systems Science*, *12*(2), 69–87. doi:10.4018/IJKSS.2021040105

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning*, 6105-6114.

Tangjaturonrasme, N., Vatanasapt, P., & Bychkov, A. (2018). Epidemiology of head and neck cancer in Thailand. *Asia Pacific Journal of Clinical Oncology*, *14*(1), 16–22. doi:10.1111/ajco.12757 PMID:28816028

Torrey, L., & Shavlik, J. (2010). Transfer learning. IGI Global. doi:10.4018/978-1-60566-766-9.ch011

Xu, L., Gao, J., Wang, Q., Yin, J., Yu, P., Bai, B., Pei, R., Chen, D., Yang, G., Wang, S., & Wan, M. (2020). Computer-aided diagnosis systems in diagnosing malignant thyroid nodules on ultrasonography: A systematic review and meta-analysis. *European Thyroid Journal*, *9*(4), 186–193. doi:10.1159/000504390 PMID:32903956

Yoo, Y. J., Ha, E. J., Cho, Y. J., Kim, H. L., Han, M., & Kang, S. Y. (2018). Computer-aided diagnosis of thyroid nodules via ultrasonography: Initial clinical experience. *Korean Journal of Radiology*, *19*(4), 665–672. doi:10.3348/kjr.2018.19.4.665 PMID:29962872

Piyanuch Arunrukthavon is currently a diagnostic radiologist at Rachapiphat Hospital. She is board certified in diagnostic radiology. She received her MD and completed her residency in diagnostic radiology from faculty of Medicine Siriraj Hospital, Mahidol University. She is interested in using innovative artificial intelligence and machine-learning to improve diagnostic imaging. She believed that further research and discoveries in this field could improve the future of diagnostic radiology.

Dittapong Songsaeng received the MD degree from Chiang Mai University, Thailand. He got a dual master's degree, M.Sc. in Neurovascular disease from Paris-Sud and Mahidol Universities, in France, and Thailand, Research fellowship in Diagnostic Neuroradiology from Massachusetts General hospital, Harvard University and Clinical fellowship from University of Toronto respectively. Currently, he is a Diagnostic and Interventional Neuroradiologist Staff and Associate Professor of Radiology at the Department of Radiology, Siriraj Hospital, Faculty of Medicine, Mahidol University, Bangkok, Thailand. His research interests lie in stroke, cerebrovascular disease, disease of thyroid gland and artificial intelligence in medicine including medical image processing, as well as, medical devices.

Chadaporn Keatmanee earned her Ph.D. in Information Science from Japan Advanced Institute of Technology and Science, and Ph.D. in Engineering and Technology from Sirindhorn International Institute of Technology, Thammasat University. She is experienced in Medical Image processing, Machine Learning, and Data Privacy and Security. As a lecturer at Ramkhamhaeng University, her goals include supporting research in medicine and encouraging students to study Computer Science.

Songphon Klabwong received the M.Sc. degree in Computer Science from the Asian Institute of Technology (AIT), Thailand. He was invited as a guest lecturer and adjunct Instructor at Thammasart University, and Bangkok University, Thailand. Currently, he is a researcher at AIT AI Center, Thailand. His research interests lie in machine learning and logic programming.

Mongkol Ekpanyapong received the B.Eng. degree in computer engineering from Chulalongkorn Univerisity, Bangkok, Thailand, in 1997, the M.Eng. degree in computer science from the Asian Institute of Technology, Thailand, in 2000, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2003 and 2006, respectively. From 1997 to 1998, he was a System Engineer with United Communication Network, Thailand. From 2006 to 2009, he was a Senior Computer Architect with Core 2 Architecture Design Team, Intel Corporation, USA. He joined the School of Engineering and Technology, Asian Institute of Technology, in 2009, where he is currently an Associate Professor. His research interests include VLSI design, physical design automation, microarchitecture, compiler, and embedded systems.

Mathew N. Dailey received the Ph.D. degree in computer science and cognitive science from the University of California, San Diego, CA, USA, in 2002. In 2006, he joined the Department of Computer Science and Information Management, Asian Institute of Technology, Bangkok, Thailand, where he is currently a professor. His research interests lie in machine learning, machine vision, robotics, and cloud computing.