Convolutional Neural Network to Stratify the Malignancy Risk of Thyroid Nodules: Diagnostic Performance Compared with the American College of Radiology Thyroid Imaging Reporting and Data System Implemented by Experienced Radiologists

G.R. Kim, E. Lee, H.R. Kim, J.H. Yoon, V.Y. Park and J.Y. Kwak

AJNR Am J Neuroradiol 2021, 42 (8) 1513-1519
doi: https://doi.org/10.3174/ajnr.A7149
http://www.ajnr.org/content/42/8/1513

Convolutional Neural Network to Stratify the Malignancy Risk of Thyroid Nodules: Diagnostic Performance Compared with the American College of Radiology Thyroid Imaging Reporting and Data System Implemented by Experienced Radiologists

G.R. Kim, E. Lee, H.R. Kim, J.H. Yoon, V.Y. Park, and J.Y. Kwak

ABSTRACT

BACKGROUND AND PURPOSE: Comparison of the diagnostic performance for thyroid cancer on ultrasound between a convolutional neural network and visual assessment by radiologists has been inconsistent. Thus, we aimed to evaluate the diagnostic performance of the convolutional neural network compared with the American College of Radiology Thyroid Imaging Reporting and Data System (TI-RADS) for the diagnosis of thyroid cancer using ultrasound images.

MATERIALS AND METHODS: From March 2019 to September 2019, seven hundred sixty thyroid nodules (≥10 mm) in 757 patients were diagnosed as benign or malignant through fine-needle aspiration, core needle biopsy, or an operation. Experienced radiologists assessed the sonographic descriptors of the nodules, and 1 of 5 American College of Radiology TI-RADS categories was assigned. The convolutional neural network provided malignancy risk percentages for nodules based on sonographic images. Sensitivity, specificity, accuracy, positive predictive value, and negative predictive value were calculated with cutoff values using the Youden index and compared between the convolutional neural network and the American College of Radiology TI-RADS. Areas under the receiver operating characteristic curve were also compared.

RESULTS: Of 760 nodules, 176 (23.2%) were malignant. At an optimal threshold derived from the Youden index, sensitivity and negative predictive values were higher with the convolutional neural network than with the American College of Radiology TI-RADS (81.8% versus 73.9%, $P = .009$; 94.0% versus 92.2%, $P = .046$). Specificity, accuracy, and positive predictive values were lower with the convolutional neural network than with the American College of Radiology TI-RADS (86.1% versus 93.7%, $P < .001$; 85.1% versus 89.1%, $P = .003$; and 64.0% versus 77.8%, $P < .001$). The area under the curve of the convolutional neural network was higher than that of the American College of Radiology TI-RADS (0.917 versus 0.891, $P = .017$).

CONCLUSIONS: The convolutional neural network provided diagnostic performance comparable with that of the American College of Radiology TI-RADS categories assigned by experienced radiologists.

ABBREVIATIONS: ACR = American College of Radiology; AUC = area under the curve; AI = artificial intelligence; CNB = core needle biopsy; CNN = convolutional neural network; FNA = fine-needle aspiration; ROC = receiver operating characteristic; TI-RADS = Thyroid Imaging and Reporting and Data System; TR = category of TI-RADS; US = ultrasound

Thyroid ultrasound (US) is the best tool to evaluate thyroid nodules for ultrasound-guided fine-needle aspiration (US-FNA).[1,2] However, the diagnostic performance of US varies because it is operator-dependent, and interobserver variability is inevitable.[3,4] To overcome this limitation, studies have been conducted on the computerized diagnosis of thyroid cancer with US images.[5-8] The convolutional neural network (CNN) is a deep learning technique that incorporates fully trainable models and can potentially cover various medical imaging tasks.[9] Recently, multiple CNN models have been investigated for the diagnosis of thyroid cancer.[10-17] Computerized algorithms were designed to

Received August 5, 2020; accepted after revision March 6, 2021.

From the Department of Radiology (G.R.K., J.H.Y., V.Y.P., J.Y.K.), Severance Hospital, Research Institute of Radiological Science, Center for Clinical Imaging Data Science, and Biostatistics Collaboration Unit (H.R.K.), Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Korea; and Department of Computational Science and Engineering (E.L.), Yonsei University, Seoul, Korea.

This study was supported by the National Research Foundation of Korea grant funded by the Korean government (Ministry of Science and ICT) (2019R1A2C1002375) and a CMB-Yuhan research grant of Yonsei University College of Medicine (6-2017-0170).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Please address correspondence to Jin Young Kwak, MD, PhD, Department of Radiology, Research Institute of Radiological Science, Center for Clinical Imaging Data Science, Yonsei University College of Medicine, 50 Yonsei-ro, Seodaemun-gu, Seoul, Republic of Korea, 120-752; e-mail: docjin@yuhs.ac

http://dx.doi.org/10.3174/ajnr.A7149

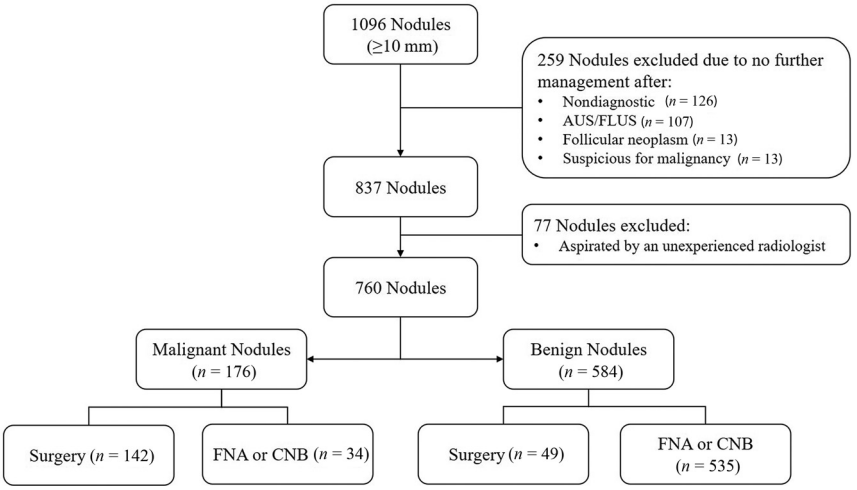

FIG 1. Flowchart of the study population. AUS/FLUS indicates atypia of undetermined significance/follicular lesion of undetermined significance.

predict thyroid cancer, and the deep CNN was used to differentiate malignant and benign thyroid nodules on the basis of US images.

Findings of past studies have been inconsistent when the diagnostic performance of the CNN was compared with visual assessment by radiologists. Even when US images were assessed according to published guidelines, the diagnostic performance of the CNN could be inferior to or favorable compared with that of radiologists, and in some studies even superior.[10-12,15] This variation might be due to unpredictable human judgment as well as differing algorithms that were developed by researchers or corporations individually; radiologists have been known to make their own final assessment, with guidelines being simply a point of reference. Thus, we aimed to compare the diagnostic performance of a CNN with a well-established guideline, the American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS), which reduces benign FNAs with high specificity and accuracy in an era when the overdiagnosis and overtreatment of thyroid cancer have become issues of concern.[18-22] ACR TI-RADS guides the diagnosis of thyroid cancer through a summation of points assigned to each US feature and then classifies nodules into 5 categories, TI-RADS (TR) 1 to TR5.[23] In our institution, the radiologist performing the US prospectively records the US features of all thyroid nodules expected to undergo US-FNA or US-guided core needle biopsy (US-CNB), and each thyroid nodule is assigned to 1 of the 5 ACR TI-RADS categories, TR1 to TR5, according to the recorded US features.

Therefore, the aim of this study was to evaluate the diagnostic performance of the CNN compared with ACR TI-RADS for the diagnosis of thyroid cancer using US images.

MATERIALS AND METHODS

Study Population

From March 2019 to September 2019, US-FNA or US-CNB was initially performed on 1096 thyroid nodules measuring ≥10 mm in 1087 patients 19 years of age or older in Severance Hospital. Of the original 1096 nodules, 259 were excluded because they did not receive further management such as repeat FNA or an operation after US-FNA showed the results as nondiagnostic ($n = 125$ in FNA; $n = 1$ in CNB). Exclusions were also due to atypia of undetermined significance/follicular lesion of undetermined significance ($n = 107$ in FNA), indeterminate ($n = 6$ in CNB), follicular neoplasm ($n = 3$ in FNA; $n = 4$ in CNB), or suspicion for malignancy ($n = 13$). Seventy-seven nodules were also excluded because they were aspirated by an inexperienced radiologist who had <1 year of experience dedicated to thyroid imaging. The remaining 760 nodules met 1 of the following criteria: 1) nodules with benign or malignant results on US-FNA or US-CNB ($n = 551$), 2) nodules that underwent an operation ($n = 191$), and 3) nodules that were confirmed as benign on repeat US-FNA or US-CNB after initial cytology results of nondiagnostic ($n = 4$) or atypia of undetermined significance/follicular lesion of undetermined significance ($n = 14$). Finally, 760 thyroid nodules in 757 patients were included (Fig 1). Three patients had 2 nodules that were aspirated from both sides of the thyroid gland.

US Image Acquisition

All US examinations were performed using a 7- to 17-mHz linear transducer (EPIQ 7; Phillips Healthcare). One of 5 radiologists dedicated to thyroid imaging with 6–21 years of experience performed the US examinations and subsequent US-FNAs. The radiologist who performed the US-FNA prospectively recorded the US features of each thyroid nodule with respect to composition, echogenicity, shape, margin, and calcifications.[2,24] Composition was assessed as solid, predominantly solid (solid component ≥50%), or predominantly cystic (solid component <50%) or spongiform. Echogenicity was assessed as hyperechoic (hyperechogenicity compared with the surrounding thyroid parenchyma), isoechoic (isoechogenicity compared with the surrounding thyroid parenchyma), hypoechoic (hypoechogenicity compared with the surrounding thyroid parenchyma), or markedly hypoechoic (hypoechogenicity compared with the strap muscles). Shape was assessed as parallel or nonparallel (greater in the anteroposterior dimension than the transverse dimension, taller-than-wide). Margin was assessed as well-defined, microlobulated, or irregular. Calcifications were classified as eggshell calcifications, macrocalcifications, microcalcifications, mixed calcifications, or no calcifications.

Image Analyses

A representative US image of each thyroid nodule was selected by an experienced radiologist (J.Y.K. with 18 years of experience dedicated to thyroid imaging), and the chosen images were stored as JPEG images in the PACS. The radiologist (J.Y.K.) drew a square ROI to cover the targeted thyroid nodule entirely using the Windows 10 Paint program. The extracted ROIs were analyzed by

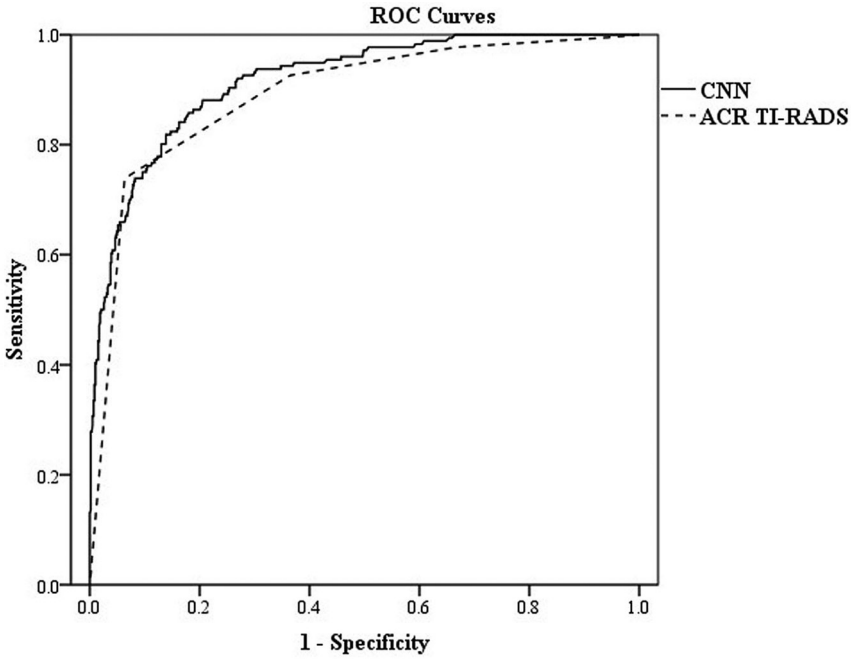

FIG 2. Comparison of ROC curves between the CNN (*solid line*) and ACR TI-RADS categories (*dotted line*). The area under the ROC curve of the CNN (0.917; 95% confidence interval, 0.895–0.936) was higher than that in the ACR TI-RADS categories (0.891; 95% confidence interval, 0.867–0.912) ($P = .017$). The areas under the ROC curve of the CNN using a malignancy risk percentage between 0 and 100 and ACR TI-RADS categories using a TR category from 1 to 5 were compared as continuous values.

the deep CNN, and malignancy risk was shown as a percentage between 0 and 100 for each thyroid nodule (Fig 2). The deep CNN implementation was based on an algorithm that was trained (fine-tuned) with 589 thyroid nodule datasets from our institution.[10] Using 3 pretrained CNNs, AlexNet, GoogLeNet, and InceptionResNetV2, we created thyroid classifiers and collected the area under the receiver operating characteristic (ROC) curve (AUC) corresponding to each CNN using Matlab 2019a (MathWorks). These classifiers and AUCs were then used to produce the mean of classification scores expressed as posterior probability in which the AUCs were used as weights. This process yields more objective results by gathering various opinions and tends to hold the final result if predictions are the same and follows the higher score if predictions contradict (see more details in the previous studies).[10,25]

One radiologist (G.R.K.) with 7 years of experience dedicated to thyroid imaging arranged the previously recorded US features to match the US descriptors used in the ACR TI-RADS and summed up the score of each nodule as follows: TR1 (0–1 point), TR2 (2 points), TR3 (3 points), TR4 (4–6 points), and TR5 (≥7 points).[21] Regarding the US features of ACR TI-RADS, "predominantly cystic" nodules were considered to have cystic or almost completely cystic composition, and "predominantly solid" nodules were considered to have mixed cystic and solid composition. "Solid" nodules were considered to have solid or almost completely solid composition. An echogenicity of "marked hypoechoic" was regarded as "very hypoechoic." "Well-defined" margins were regarded as smooth and microlobulated, and "irregular" margins were regarded as lobulated or irregular. "Eggshell calcifications" were regarded as peripheral (rim) calcifications. and "mixed calcification" and "microcalcifications" were regarded as punctate echogenic.

Data and Statistical Analysis

Benign results on US-FNA or US-CNB and benign or malignant histopathologic results from an operation and follow-up US-FNA or US-CNB were the reference standards for analysis. On the basis of these results, we calculated the malignancy risk of the 5 categories of ACR TI-RADS, respectively. Each nodule that had its percentage of malignancy risk calculated by the CNN was re-categorized into 1 of the 5 TR categories according to the malignancy risk range suggested for each TR category by ACR TI-RADS.[22,26] Malignancy risk was also calculated for those TR categories created from the CNN (CNN-TR).

Variables were compared between the benign and malignant nodules using the Mann-Whitney *U* test and the χ^2 test or the Fisher exact test. Diagnostic performances including sensitivity, specificity, accuracy, positive predictive value, and negative predictive value for predicting thyroid malignancy were calculated for the CNN and ACR TI-RADS with 95% confidence intervals. The cutoff value to diagnose thyroid malignancy was defined using the Youden index in the CNN (malignancy risk percentage as a continuous variable) and ACR TI-RADS (TR category as an ordinal variable).[27] Logistic regression using the generalized estimating equation method was used to test the significance of comparisons with adjustments for correlated observations of clustered data. The AUCs of the CNN using a malignancy risk percentage between 0 and 100 and ACR TI-RADS categories using a TR category from 1 to 5 were compared as continuous values using the DeLong method.[28]

All statistical analyses were performed with SAS (Version 9.4; SAS Institute) and SPSS 25.0 for Windows (IBM). Statistical significance was defined with P values $< .05$.

RESULTS

Study Population and Nodule Characteristics

In 760 thyroid nodules, 176 (23.2%) were malignant. Final diagnoses of the 176 malignant nodules were confirmed through surgical resection ($n = 142$; one hundred thirty-two papillary thyroid carcinomas, 5 follicular carcinomas, 2 poorly differentiated carcinomas, 1 medullary carcinoma, 1 Hurthle cell carcinoma, and 1 squamous cell carcinoma) and US-FNA ($n = 34$; 33 papillary thyroid carcinomas and 1 small-cell carcinoma). The median size of all 176 nodules was 20 mm (interquartile range, 14–30 mm). The median age of the 757 patients was 51 years (interquartile range,

Table 1: Patient demographics and distribution of ACR TI-RADS features in benign and malignant thyroid nodules ($n = 760$)[a]

Characteristics	All ($n = 760$)	Benign Nodules ($n = 584$)	Malignant Nodules ($n = 176$)	*P* Value
Sex				.035
Women	587	462 (79.4%)	125 (71.4%)	
Men	170	120 (20.6%)	50 (28.6%)	
Age (median) (interquartile range) (yr)	51 (39–61)	52 (41–61)	45 (34–60)	<.001
Nodule size (median) (interquartile range) (mm)	20 (14–30)	23 (15–32)	14 (11–20)	<.001
Nodule features				
Composition				<.001
Cystic or almost completely cystic	50	47 (8.0%)	3 (1.7%)	
Spongiform	1	1 (0.2%)	0	
Mixed cystic and solid	234	222 (38.0%)	12 (6.8%)	
Solid or almost completely solid	475	314 (53.8%)	161 (91.5%)	
Echogenicity				<.001
Anechoic		0	0	
Hyperechoic or isoechoic	410	387 (66.3%)	23 (13.1%)	
Hypoechoic	329	191 (32.7%)	138 (78.4%)	
Very hypoechoic	21	6 (1.0%)	15 (8.5%)	
Shape				<.001
Wider-than-tall	671	554 (94.9%)	117 (66.5%)	
Taller-than-wide	89	30 (5.1%)	59 (33.5%)	
Margin				<.001
Smooth	579	535 (91.6%)	44 (25.0%)	
Ill-defined	0	0	0	
Lobulated or irregular	181	49 (8.4%)	132 (75.0%)	
Extrathyroidal extension	0	0	0	
Echogenic foci				<.001
None or large comet-tail artifacts	536	477 (81.7%)	59 (33.5%)	
Macrocalcifications	91	69 (11.8%)	22 (12.5%)	
Peripheral (rim) calcifications	10	10 (1.7%)	0	
Punctate echogenic foci	123	28 (4.8%)	95 (54.0%)	

[a] Data are numbers of nodules, with percentages in parentheses.

Table 2: Calculated malignancy risk of each category according to the risk stratification of ACR TI-RADS

	TR1	TR2	TR3	TR4	TR5	Total
Suggested risk of malignancy (%)[20,24]	≤2	≤2	2< and ≤5	5< and ≤20	>20	
ACR TI-RADS category						
No. of malignant nodules	0	4	9	33	130	176
Assigned total nodules	41	158	185	209	167	760
Calculated risk of malignancy (%)	0	2.5	4.9	15.8	77.8	23.2
CNN[a]						
No. of malignant nodules	0	0	0	9	167	176
Assigned total nodules	0	5	45	307	403	760
Calculated risk of malignancy (%)	0	0	0	2.9	41.4	23.2

[a] Malignancy percentages provided by the CNN were re-categorized according to the suggested cancer risk levels of ACR TI-RADS.

39–61 years). Of the 757 patients, 587 (77.5%) were women and 170 (22.5%) were men.

The US features of the benign and malignant nodules according to ACR TI-RADS and their distributions are described in Table 1. The median size of the benign nodules was 23 mm, which was larger than the that of malignant nodules (median, 14 mm; $P < .001$). Solid or almost completely solid composition (161 of 176, 91.5%), hypoechoic or very hypoechoic echogenicity (153 of 176, 86.9%), taller-than-wide shape (59 of 176, 33.5%), lobulated or irregular margins (132 of 176, 75.0%), and punctate echogenic foci (95 of 176, 54.0%) were frequently seen in the malignant nodules ($P < .001$, respectively).

Malignancy Risk According to ACR TI-RADS Category

Table 2 summarizes the malignancy risk of each category in ACR TI-RADS and CNN-TR that was calculated after nodules were re-categorized according to the malignancy-risk ranges suggested by ACR TI-RADS.[22,26] The malignancy risk of ACR TR5 was 77.8% (130 of 167), which was much higher than the suggested malignancy risk of 20%. The malignancy risks of ACR TR1 to TR4 were within the risk ranges suggested by the ACR. According to the CNN, 403 thyroid nodules had malignancy risks higher than 20% and were re-categorized as CNN-TR5. Among 403 nodules, 167 were thyroid cancers (41.4%). Of 760 nodules, 307 nodules that had a 5%–20% range of malignancy risk according to the CNN were re-categorized to CNN-TR4 and 9 of these 307 (2.9%) nodules were thyroid cancers.

Comparing the Diagnostic Performances of CNN and ACR TI-RADS

According to the cutoff value found using the Youden index in the CNN and ACR TR categories, respectively, thyroid nodules

Table 3: Comparison of diagnostic performance between CNN and ACR TI-RADS

	CNN (95% CI)	ACR TI-RADS (95% CI)	*P* Value
Sensitivity	81.8% (76.1–87.5)	73.9% (67.4–80.4)	.009
Specificity	86.1% (83.3–88.9)	93.7% (91.7–95.6)	<.001
Accuracy	85.1% (82.6–87.7)	89.1% (86.9–91.3)	.003
Positive predictive value	64.0% (57.7–70.3)	77.8% (71.6–84.1)	<.001
Negative predictive value	94.0% (92–96)	92.2% (90.1–94.4)	.046
AUC[a]	0.917 (0.895-0.936)	0.891 (0.867-0.912)	.017

[a] The AUCs of the CNN using a malignancy risk percentage between 0 and 100 and ACR TI-RADS categories using a TR category from 1 to 5 were compared as continuous values.

with a malignancy risk of 52.6% or higher in the CNN or nodules equal to or higher than TR5 according to ACR TI-RADS were considered malignant. The diagnostic performances of the ACR TI-RADS and CNN are summarized in Table 3. Sensitivity was significantly higher with the CNN than with ACR TI-RADS (81.8% versus 73.9%, $P = .009$). Specificity, accuracy, and positive predictive values were significantly lower with the CNN than with ACR TI-RADS (86.1% versus 93.7%, $P < .001$; 85.1% versus 89.1%, $P = .003$; and 64.0% versus 77.8%, $P < .001$, respectively). The negative predictive value was significantly higher with the CNN than with ACR TI-RADS (94.0% versus 92.2%, $P = .046$). Figure 2 shows the ROC curves for the diagnosis of thyroid cancer with the CNN and ACR TI-RADS. The AUC of the CNN (0.917; 95% CI, 0.895–0.936) was higher than that of the ACR TI-RADS categories (0.891; 95% CI, 0.867–0.912) ($P = .017$).

DISCUSSION

Our study demonstrates that the CNN shows diagnostic performance comparable with that of ACR TI-RADS when experienced radiologists assigned US descriptors and scored their observations. The malignancy risk of each ACR TR category in our study was within the range suggested by ACR TI-RADS. In our study, the sensitivity (81.8%), specificity (86.1%), and accuracy (85.1%) of the CNN were within ranges similar to those reported in previous publications on the deep CNN for the diagnosis of thyroid cancer.[10,11,16,17] At an optimal threshold derived from the Youden index, our CNN was more sensitive but less specific and accurate compared with the ACR TI-RADS (sensitivity, 81.8% versus 73.9%; specificity, 86.1% versus 93.7%; and accuracy, 85.1% versus 89.1%). The AUC was higher in the CNN than in ACR TI-RADS (0.917 versus 0.891, $P = .017$).

Past studies have shown different results for the diagnostic performance of the CNN compared with visual assessment by radiologists. According to Ko et al,[10] the CNN showed favorable diagnostic performances for predicting thyroid cancer on US, with sensitivities of 84.0%–91.0%, specificities of 82.0%–90.0%, accuracies of 86.0%–88.0%, and AUCs of 0.835–0.850, values that were like those of experienced radiologists. Li et al[11] reported somewhat higher performances for the CNN with AUCs of 0.908–0.947; compared with experienced radiologists, the CNN showed similar sensitivity (84.3%–93.4%) and significantly higher specificity (86.1%–87.3%) and accuracy (85.7%–89.8%). Unlike the favorable performances of the CNN observed in the above-mentioned studies, the CNN in the study of Kim et al[12] had lower specificity (68.2%) and accuracy (73.4%) compared with radiologists for the diagnosis of thyroid cancer, even though it achieved similar sensitivity (81.4%). In our study, the specificity and accuracy of the CNN were somewhat higher than those reported in Kim et al. The different frequencies of punctate echogenic foci (considered as microcalcifications) in malignant nodules (54.0% in our study versus 72.1% in Kim et al) might be 1 explanation because Kim et al suggested the recognition of microcalcifications as a cause of inaccuracy for the CNN in their study. In addition, the inferior performance of the CNN in their study was thought to originate from manual manipulation for segmentation and human-designed features applied to the computer-aided diagnosis system. Moreover, the experience level of the performing operator had an effect on the performance of computer-aided diagnosis because of the manual manipulation required for computer-aided diagnosis.[29]

Unlike the traditional machine learning algorithm or the traditional commercial system that is connected to US machines and already applied in clinical practice,[5,12,29] the recently introduced deep CNN is not limited to or influenced by human-designed features known to represent thyroid cancer on US, though its operational principles for diagnosing thyroid malignancy are not yet completely explained by humans. In our study, the radiologist just drew a square ROI covering the entire targeted nodule without any human interference with the diagnostic process of the CNN. Instead of using features engineered by humans, the deep CNN extracts image information directly from imaging data, and the CNN might be able to recognize cancer-specific US features that are not identified explicitly by the naked eye.[30]

Because US is performed and interpreted by humans, any diagnosis of thyroid cancer based on US images is subjective, thus requiring experience and expertise.[3,4] Recent studies have evaluated the computer-aided diagnosis of thyroid cancer, which incorporates texture analysis and machine learning and deep learning techniques for US images; the authors reported that computer-aided diagnosis showed comparable and even higher diagnostic performance compared with radiologists.[5,6,11,29] While artificial intelligence (AI) is not yet considered ready for a clinical setting,[31] computer software is already thought to have several strong advantages over radiologists because its use can overcome human variation and provide diagnostic reproducibility and consistency in image interpretation. However, past studies have shown greatly differing results when the diagnostic performance of the CNN is compared with human interpretation. This might be due to the diversity of assessments possible by radiologists as well as the different algorithms developed by individual researchers or corporations. Despite referring to guidelines, radiologists might eventually reach diagnoses independently on the basis of their individual expertise and experience.

On the other hand, we intended to directly compare the performances of our CNN with that of an established guideline, ACR TI-RADS, which is known to have a high specificity and positive predictive value without sacrificing sensitivity, and to further use this knowledge to help radiologists achieve optimal performances with the ACR TI-RADS.[20,21,32,33] We used results found with prospectively recorded descriptors that were obtained during real-time evaluations of entire 3D nodules instead of those collected through a retrospective human review of single US images. This choice might represent ACR TI-RADS more properly and objectively than a new individual human review. The experienced radiologists in the study of Li et al[11] showed a less specific and accurate performance than the CNN; the radiologists in the study of Li et al showed low specificities of 57.1%–68.6% and low accuracies of 72.7%–78.8% compared with the previous studies and our study. Regarding this matter, Li et al replied that their reviewers were burdened due to the larger subject sample and subsequent large amounts of image reviews needed.[11,34,35]

In this study, the malignancy risk of each ACR TR category was within the theoretic percentage of malignancy risk, which meant that nodules had been assessed appropriately with ACR TI-RADS. The ACR TR categories of our study showed enough specificity and accuracy for diagnosis, fulfilling the original goals of ACR TI-RADS to decrease biopsies with benign findings and improve accuracy. On the other hand, radiologists have shown a wide range in diagnostic performance with ACR TI-RADS because sensitivity has been reported to be 81.7%–96.7%; specificity, 47.7%–77.3%; and accuracy, 69.3%–84.9%.[20,21,32,36] This inconsistency in performance might be caused by the different experience levels of the radiologists or by the different cutoff values of each study. In our study, we were able to conduct a relatively objective validation of ACR TI-RADS by experienced radiologists using US features and to compare its diagnostic performance with that of the AI diagnosis. The diagnostic performance of the CNN was comparable with that of ACR TI-RADS with a somewhat higher AUC for thyroid cancer. Given that a recent study reported that alteration of ACR TI-RADS by AI led to improvement in specificity, the adequate modification and fusion of the settled guidelines and AI, ie, AI–powered US, might be a potential aid to better diagnostic performance and implementation of AI.[36,37]

This study has several limitations. First, US examinations are performed in real-time. The process of image acquisition such as capturing 2D-US images and selecting a representative image from the acquired images is inevitably operator-dependent. Additionally, there are limits to how much 2D US images can represent the entire thyroid nodule. AI studies that analyze 3D-US images might be of more help in the future.[37] Second, we used data prospectively recorded in our institutional data base, in which US features were described with different terminology than that suggested by the ACR guidelines. Because information about "anechoic," "ill-defined" or "extrathyroidal," and large comet-tail artifacts was not collected during the study period, despite being listed in the ACR guidelines, this issue might be a limitation of our study. However, we did not conduct an intentional retrospective review for this study because we aimed to investigate ACR TI-RADS itself and not the man-made final assessments. Third, our institution is a tertiary center, and we included thyroid nodules that underwent US-FNA or US-CNB only, which meant that surgical histopathology was unavailable. Thus, there might be false-negative or false-positive results, even though the rates would be very low with a false-negative rate of <3% and a false-positive rate of about 3%–4%.[38] Fourth, the ROC-derived cutoff value that we used to calculate diagnostic performance cannot be accepted as a diagnostic standard in real clinical practice without further validation.

CONCLUSIONS

The CNN provided diagnostic performance comparable with that of the ACR TI-RADS categories assigned by experienced radiologists. Before AI can be used to diagnose thyroid cancer, a thorough evaluation of AI diagnosis compared with pre-existing guidelines is needed, and our study should be able to present a relatively objective comparison of diagnostic performances between the ACR TI-RADS and CNN for thyroid cancer. Adequate modification and fusion of the ACR TI-RADS and CNN that takes advantage of their unique characteristics will help optimize overall diagnostic performance.

Disclosures: Jin Young Kwak—*RELATED*: *Grant*: National Research Foundation of Korea grant funded by the Korean government (Ministry of Science and ICT) (2019R1A2C1002375).

REFERENCES

1. Haugen BR, Alexander EK, Bible KC, et al. **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* 2016;26:1–133 CrossRef Medline
2. Kwak JY, Han KH, Yoon JH, et al. **Thyroid imaging reporting and data system for US features of nodules: a step in establishing better stratification of cancer risk.** *Radiology* 2011;260:892–99 CrossRef Medline
3. Choi SH, Kim EK, Kwak JY, et al. **Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules.** *Thyroid* 2010;20:167–72 CrossRef Medline
4. Park SH, Kim SJ, Kim EK, et al. **Interobserver agreement in assessing the sonographic and elastographic features of malignant thyroid nodules.** *AJR Am J Roentgenol* 2009;193:W416–23 CrossRef Medline
5. Chang Y, Paul AK, Kim N, et al. **Computer-aided diagnosis for classifying benign versus malignant thyroid nodules based on ultrasound images: a comparison with radiologist-based assessments.** *Med Phys* 2016;43:554 CrossRef Medline
6. Choi YJ, Baek JH, Park HS, et al. **A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: initial clinical assessment.** *Thyroid* 2017;27:546–52 CrossRef Medline
7. Acharya UR, Swapna G, Sree SV, et al. **A review on ultrasound-based thyroid cancer tissue characterization and automated classification.** *Technol Cancer Res Treat* 2014;13:289–301 CrossRef Medline
8. Gopinath B, Shanthi N. **Computer-aided diagnosis system for classifying benign and malignant thyroid nodules in multi-stained FNAB cytological images.** *Australas Phys Eng Sci Med* 2013;36:219–30 CrossRef Medline
9. Shin HC, Roth HR, Gao M, et al. **Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning.** *IEEE Trans Med Imaging* 2016;35:1285–98 CrossRef Medline

10. Ko SY, Lee JH, Yoon JH, et al. **Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound.** *Head Neck* 2019;41:885–91 CrossRef Medline
11. Li X, Zhang S, Zhang Q, et al. **Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study.** *Lancet Oncol* 2019;20:193–201 CrossRef Medline
12. Kim HL, Ha EJ, Han M. **Real-world performance of computer-aided diagnosis system for thyroid nodules using ultrasonography.** *Ultrasound Med Biol* 2019;45:2672–78 CrossRef Medline
13. Ma J, Wu F, Zhu J, et al. **A pre-trained convolutional neural network based method for thyroid nodule diagnosis.** *Ultrasonics* 2017;73:221–30 CrossRef Medline
14. Ma J, Wu F, Jiang T, et al. **Cascade convolutional neural networks for automatic detection of thyroid nodules in ultrasound images.** *Med Phys* 2017;44:1678–91 CrossRef Medline
15. Buda M, Wildman-Tobriner B, Hoang JK, et al. **Management of thyroid nodules seen on US images: deep learning may match performance of radiologists.** *Radiology* 2019;292:695–701 CrossRef Medline
16. Koh J, Lee E, Han K, et al. **Diagnosis of thyroid nodules on ultrasonography by a deep convolutional neural network.** *Sci Rep* 2020;10:15245 CrossRef Medline
17. Jin Z, Zhu Y, Zhang S, et al. **Ultrasound computer-aided diagnosis (CAD) based on the thyroid imaging reporting and data system (TI-RADS) to distinguish benign from malignant thyroid nodules and the diagnostic performance of radiologists with different diagnostic experience.** *Med Sci Monit* 2020;26:e918452 CrossRef Medline
18. Park S, Oh CM, Cho H, et al. **Association between screening and the thyroid cancer "epidemic" in South Korea: evidence from a nationwide study.** *BMJ* 2016;355:i5745 CrossRef Medline
19. Jegerlehner S, Bulliard JL, Aujesky D, et al. NICER Working Group. **Overdiagnosis and overtreatment of thyroid cancer: a population-based temporal trend study.** *PLoS One* 2017;12:e0179387 CrossRef Medline
20. Wu XL, Du JR, Wang H, et al. **Comparison and preliminary discussion of the reasons for the differences in diagnostic performance and unnecessary FNA biopsies between the ACR TIRADS and 2015 ATA guidelines.** *Endocrine* 2019;65:121–31 CrossRef Medline
21. Yoon JH, Lee HS, Kim EK, et al. **Pattern-based vs. score-based guidelines using ultrasound features have different strengths in risk stratification of thyroid nodules.** *Eur Radiol* 2020;30:3793–3802 CrossRef Medline
22. Tappouni RR, Itri JN, McQueen TS, et al. **ACR TI-RADS: pitfalls, solutions, and future directions.** *Radiographics* 2019;39:2040–52 CrossRef Medline
23. Tessler FN, Middleton WD, Grant EG, et al. **ACR thyroid imaging, reporting and data system (TI-RADS): White Paper of the ACR TI-RADS Committee.** *J Am Coll Radiol* 2017;14:587–95 CrossRef Medline
24. Kim EK, Park CS, Chung WY, et al. **New sonographic criteria for recommending fine-needle aspiration biopsy of nonpalpable solid nodules of the thyroid.** *AJR Am J Roentgenol* 2002;178:687–91 CrossRef Medline
25. Lee E, Ha H, Kim HJ, et al. **Differentiation of thyroid nodules on US using features learned and extracted from various convolutional neural networks.** *Sci Rep* 2019;9:19854 CrossRef Medline
26. Middleton WD, Teefey SA, Reading CC, et al. **Comparison of performance characteristics of American College of Radiology TI-RADS, Korean Society of Thyroid Radiology TIRADS, and American Thyroid Association Guidelines.** *AJR Am J Roentgenol* 2018;210:1148–54 CrossRef Medline
27. Youden WJ. **Index for rating diagnostic tests.** *Cancer* 1950;3:32–35 CrossRef Medline
28. DeLong ER, DeLong DM, Clarke-Pearson DL. **Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.** *Biometrics* 1988;44:837–45 Medline
29. Jeong EY, Kim HL, Ha EJ, et al. **Computer-aided diagnosis system for thyroid nodules on ultrasonography: diagnostic performance and reproducibility based on the experience level of operators.** *Eur Radiol* 2019;29:1978–85 CrossRef Medline
30. LeCun Y, Bengio Y, Hinton G. **Deep learning.** *Nature* 2015;521:436–44 CrossRef Medline
31. Park SH, Han K. **Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction.** *Radiology* 2018;286:800–09 CrossRef Medline
32. Ruan JL, Yang HY, Liu RB, et al. **Fine needle aspiration biopsy indications for thyroid nodules: compare a point-based risk stratification system with a pattern-based risk stratification system.** *Eur Radiol* 2019;29:4871–78 CrossRef Medline
33. Ha EJ, Na DG, Baek JH, et al. **US fine-needle aspiration biopsy for thyroid malignancy: diagnostic performance of seven society guidelines applied to 2000 thyroid nodules**. *Radiology* 2018;287:893–900 CrossRef Medline
34. Ha EJ, Baek JH, Na DG. **Deep convolutional neural network models for the diagnosis of thyroid cancer.** *Lancet Oncol* 2019;20:e130 e130 CrossRef Medline
35. Li X, Zhang S, Zhang Q, et al. **Deep convolutional neural network models for the diagnosis of thyroid cancer: authors' reply.** *Lancet Oncol* 2019;20:e131 CrossRef Medline
36. Wildman-Tobriner B, Buda M, Hoang JK, et al. **Using artificial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: diagnostic accuracy and utility.** *Radiology* 2019;292:112–19 CrossRef Medline
37. Akkus Z, Cai J, Boonrod A, et al. **A survey of deep-learning applications in ultrasound: artificial intelligence-powered ultrasound for improving clinical workflow.** *J Am Coll Radiol* 2019;16:1318–28 CrossRef Medline
38. Cibas ES, Ali SZ. **The 2017 Bethesda System for reporting thyroid cytopathology.** *Thyroid* 2017;27:1341–46 CrossRef Medline