scientific reports

OPEN

Diagnosis of thyroid micronodules on ultrasound using a deep convolutional neural network

Miribi Rho[1], Sei Hyun Chun[1], Eunjung Lee[2], Hye Sun Lee[3], Jung Hyun Yoon[1], Vivian Youngjean Park[1], Kyunghwa Han[1] & Jin Young Kwak[1]✉

To assess the performance of deep convolutional neural network (CNN) to discriminate malignant and benign thyroid nodules < 10 mm in size and compare the diagnostic performance of CNN with those of radiologists. Computer-aided diagnosis was implemented with CNN and trained using ultrasound (US) images of 13,560 nodules ≥ 10 mm in size. Between March 2016 and February 2018, US images of nodules < 10 mm were retrospectively collected at the same institution. All nodules were confirmed as malignant or benign from aspirate cytology or surgical histology. Diagnostic performances of CNN and radiologists were assessed and compared for area under curve (AUC), sensitivity, specificity, accuracy, positive predictive value, and negative predictive value. Subgroup analyses were performed based on nodule size with a cut-off value of 5 mm. Categorization performances of CNN and radiologists were also compared. A total of 370 nodules from 362 consecutive patients were assessed. CNN showed higher negative predictive value (35.3% vs. 22.6%, P = 0.048) and AUC (0.66 vs. 0.57, P = 0.04) than radiologists. CNN also showed better categorization performance than radiologists. In the subgroup of nodules ≤ 5 mm, CNN showed higher AUC (0.63 vs. 0.51, P = 0.08) and specificity (68.2% vs. 9.1%, P < 0.001) than radiologists. Convolutional neural network trained with thyroid nodules ≥ 10 mm in size showed overall better diagnostic performance than radiologists in the diagnosis and categorization of thyroid nodules < 10 mm, especially in nodules ≤ 5 mm.

Abbreviations

US	Ultrasound
FNA	Fine-needle aspiration
CNN	Convolutional neural network
TIRADS	Thyroid Imaging Reporting and Data System
KSThR	Korean Society of Thyroid Radiology
CAD	Computer-aided diagnosis
AUC	Area under the receiver characteristic curve

The detection of thyroid nodules has substantially increased with the widespread use of high-resolution ultrasound (US), resulting in a high prevalence of 19–67% for thyroid nodules in the general population[1,2]. Approximately 7–15% of detected thyroid nodules are thyroid cancers[3]. In thyroid micronodules (< 10 mm), fine-needle aspiration (FNA) remains controversial because papillary thyroid microcarcinomas, defined as tumors < 10 mm in size, have shown near-zero cancer-specific mortality[4]. As it is difficult to predict which thyroid microcarcinoma will progress with clinical significance, most guidelines simply state FNA as an available option, leaving the decision up to clinicians to decide based on clinical settings and patient preference[5–8].

Multifocality and bilaterality in papillary thyroid carcinoma are common features with a reported frequency of 18–87%[9], and are known risk factors of nodal metastasis, distant metastasis, and regional recurrence after initial therapy[10]. The American Thyroid Association guideline first recommends lobectomy for unifocal papillary thyroid microcarcinoma without extrathyroidal extension but also notes that the presence of a bilateral nodule can suggest the need for a bilateral thyroidectomy to address the possibility of bilaterality[5]. To note, the US features used to differentiate benign and malignant thyroid nodules are equally applied to both macronodules

[1]Department of Radiology, Severance Hospital, Research Institute of Radiological Science, Yonsei University College of Medicine, Seoul, Korea. [2]School of Mathematics and Computing, Yonsei University, Seoul, Korea. [3]Biostatistics Collaboration Unit, Yonsei University College of Medicine, Seoul, Korea. ✉email: docjin@yuhs.ac

and micronodules)[11,12]. Considering that physicians' visual analysis of micronodules on US, especially of nodules smaller than 5 mm, has shown high false-positive rates, the preoperative detection of micronodules may increase additional FNA[13,14]. Furthermore, given the high nondiagnostic rate of FNA, preoperative diagnosis is still a challenging task for micronodules[10,14].

The convolutional neural network (CNN) is a deep learning model which enables high-performance visual recognition and classification after automatically learning representative features from a training set[15,16]. The characteristics of the training set are therefore critical to the performance of CNN. CNN-based methods have been investigated to differentiate malignant and benign thyroid nodules and showed non-inferior or comparable diagnostic performance to radiologists[17–25]. Most studies have been conducted on thyroid nodules over 10 mm, and only three included thyroid nodules larger than 5 mm, but their mean size was larger than 10 mm[20,24,25]. Three other investigations have shown validation results for nodules corresponding to the same size criteria with training sets made up of nodules larger than 10 mm[18,21] or 5 mm[20], while no other study has demonstrated nodule size criteria in both the training and validation of CNN[17,19,22–25]. To the best of our knowledge, no study has applied a CNN-based model to thyroid nodules beyond the size criteria of the training set. In this study, we investigated the diagnostic performances of a CNN that was previously trained with thyroid nodules ≥ 10 mm to discriminate malignant and benign thyroid nodules < 10 mm and compare its diagnostic performance with those of radiologists.

Methods

The institutional review board of Severance Hospital (Seoul, South Korea) approved this retrospective study, with a waiver for informed consent (IRB number: 2020-3659-001). Signed informed consent for biopsy or surgical procedures was obtained preoperatively from all patients. All methods were performed in accordance with relevant guidelines and regulations.

Patients.

This study was performed at a single tertiary referral center from March 2016 to February 2018, during which 4110 nodules in 3716 consecutive patients were consulted for US-guided FNA. The initial FNA was performed in 3323 nodules in 3240 patients, of which 698 nodules were < 10 mm in 683 patients. Our study included nodules < 10 mm if they (a) were cytologically confirmed as benign or malignant (Bethesda category II or VI) or (b) were confirmed as malignant on postsurgical histology. We excluded nodules that were not confirmed or lost to follow-up. Finally, a total of 370 thyroid nodules in 362 patients were included and analyzed (Fig. 1). Two thyroid nodules were included for 8 patients, among which 6 patients had both malignant nodules and 2 patients had one benign and one malignant nodule.

US imaging.

US examinations of both thyroid glands and neck areas were performed using a 5–12 MHz linear array transducer (*i*U22, Philips Healthcare, Amsterdam, Netherlands). Real-time US scans and subsequent US-FNA were performed by 12 radiologists with 1–20 years of experience in thyroid imaging.

Each radiologist who performed the US and US-FNA/core biopsy procedures interpreted each US scan of the thyroid nodules and recorded US features prospectively in our institutional database[26,27]. US features including composition, echogenicity, margin, calcifications, and shape were recorded using descriptors that have been used from June 2012 to the present in our institution[28]. Each thyroid nodule was categorized according to the Thyroid

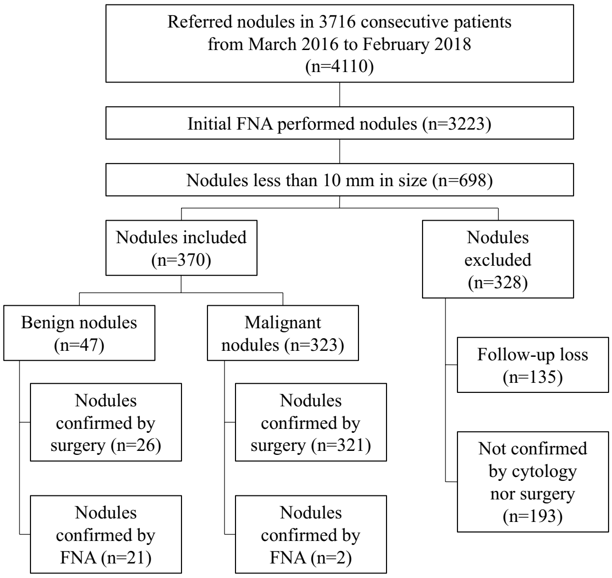

Figure 1. Flow chart of patient enrollment. A total of 370 nodules including 322 malignant nodules and 48 benign nodules were included in this study. *FNA* fine-needle aspiration.

Imaging Reporting and Data System suggested by the Korean Society of Thyroid Radiology (KSThR TIRADS) using pre-recorded US features[7].

Image acquisition and CNN evaluation. An experienced radiologist with 20 years of experience dedicated to thyroid imaging who was blinded to clinical information and pathological results selected and retrieved a representative US image for each thyroid nodule from the PACS and stored it in JPEG format. For each image, a square ROI enclosing the entire targeted thyroid nodule was manually labeled using the Paint program of Windows 10 by the same radiologist who retrieved the images.

We used a computer-aided diagnosis (CAD) program to assess the malignancy risk of 370 thyroid nodules on US images. The performance of a CNN algorithm differs by data set, that is, it highly depends on the data used to train its network. There are many pre-trained models and a few of their test results (accuracy, sensitivity, and specificity of 370 test data sets) are reported in Supplemental Table S1. As ResNet101 shows one of the best performances with current US images, this paper focuses on analyzing the results from transfer learning using ResNet101. The pretrained CNN model ResNet101[29,30] was fine-tuned with 13,560 US images of thyroid nodules ≥ 10 mm in size (further details on the CAD program are provided in the Supplemental Material)[21]. ResNet101 is a deep neural network that was originally trained with 1000 object classes, 1,281,167 training images, and 50,000 validation images. The basic algorithm of the residual net family (ResNet-18,34,50,101, and 152) has been previously introduced[29] and the paper achieved state-of-the-art results in image classification by taking a standard feed-forward ConvNet and adding skip-connections that bypassed a few convolution layers at a time. Each bypass/shortcut produced a residual block from which the convolution layers predicted a residual further used in the block's input tensor. ResNet101 consists of 347 layers capable of learning rich feature representations of images with an image input size of 224-by-224. For transfer learning, 13,560 US images composed of 7160 malignant and 6400 benign nodule images were used. To balance the number of data sets, we used the left–right mirroring augmentation of 760 randomly selected benign images so that a final total of 14,320 images were used in training. Since the fully connected layer and classification layer at the end of the original pretrained network were configured for 1000 classes, they were replaced with new layers adapted to the new data set (benign and malignant) with learning rates for weights and biases set to 10 each. In the fine-tuning process, the stochastic gradient descent with a momentum optimizer was used to train the network, the initial learning rate was set to 10-4, 10 epochs were conducted, and the mini-batch size was set to 50. The momentum of the stochastic gradient descent optimizer was set to 0.9 and the learning rate dropped by a factor of 0.5 every 4 epochs. The model was validated with internal data (95 benign, 539 malignant) and external data from three different hospitals (429 benign, 761 malignant).

Using the CAD program, we calculated the risks of malignancy as continuous values ranging from 0 to 100% (CAD value). We also categorized nodules by designating categories based on the CAD value (CNN TIRADS) according to the predicted probability from KSThR TIRADS. CNN TIRADS category 2 was assigned to nodules with a malignancy probability < 3%, category 3 for a probability < 15%, category 4 for a probability < 60% and category 5 for a probability ≥ 60%[7].

Statistical analysis. For the reference standard, histopathologic results from FNA or surgery were used to confirm the final diagnosis of each thyroid nodule. If there was a discrepancy between the two results, the reference standard was the histopathologic result from the surgical specimen.

Baseline patient characteristics and nodal US features were compared between malignant and benign nodules with the Student's *t*-test and Pearson's χ^2-test at the patient level and the logistic regression analysis with the generalized estimating equation method for clustered data in a nodule-level comparison. Areas under the receiver operating characteristics curve (AUCs) with 95% CIs were obtained and the TIRADS category and CAD value of each thyroid nodule were divided as either positive or negative according to the Youden index. We compared the diagnostic performances of the TIRADS category and CNN by analyzing the sensitivity, specificity, accuracy, positive predictive value, and negative predictive value using logistic regression with the generalized estimating equation method. AUC values were compared with the Obuchowski algorithm for clustered data[31]. The same statistical analysis was performed for the subgroup analysis separately according to nodule size with a cut-off value of 5 mm.

We assessed the categorization performances of CNN TIRADS and KSThR TIRADS using the likelihood ratio χ^2-test and the linear trend χ^2-test for each categorization system to determine heterogeneity (small differences in risk of malignancy among nodules in the same category) and monotonicity of gradients (whether the risk of malignancy of nodules increases as the category increases), respectively[32,33]. We also used the Akaike information criterion, which is a widely used estimator for model selection. Smaller Akaike information criterion values indicate a more informative model in terms of goodness of fit[34].

Statistical analysis was performed using statistical software (SAS version 9.4, SAS Institute, Cary, NC, USA) and the R Statistical Package (Version 4.0.2, Institute for Statistics and Mathematics, Vienna, Austria). Two-sided *P* values < 0.05 were considered to indicate statistical significance.

Results

Patients and nodules characteristics. A total of 370 nodules in 362 patients (mean ages, 46 ± 12 years; range 20–76 years) made up of 289 (79.8%; mean ages, 46 ± 12 years; range, 20–76 years) women and 73 (20.2%; mean ages, 45 ± 12 years; range, 26–73 years) men, were included in the final study population (Fig. 1). There were 347 (93.8%) nodules which were confirmed with surgery and 23 (6.2%) nodules which were confirmed with FNA. FNA was performed in the 370 nodules because of requests from physicians at outside clinics (n = 127), high suspicion nodules > 5 mm (n = 123)[7], the need to determine surgical extent in patients with bilateral nodules (n = 83), patient request (n = 30) and cervical lymph node metastasis (n = 7).

Among the 370 nodules, 323 nodules were confirmed as malignant and 47 nodules were confirmed as benign (Figs. 2 and 3). Of these malignant nodules, 322 nodules were confirmed as papillary thyroid carcinoma and 1 nodule as medullary thyroid carcinoma. The mean nodule size of the malignant and benign nodules was 5.3 ± 1.5 mm and 5.8 ± 2.2 mm, respectively (P = 0.14, Table 1). No significant difference was observed between the malignant and benign nodules for age (46.0 years vs. 45.9 years, P = 0.97) and female proportion (79.2% vs. 85.1%, P = 0.34).

Comparison of diagnostic performance between the CNN and radiologists. The optimal cut-off points, set by the Youden index, were probability > 56.1% for CNN and KSThR TIRADS category 5 for radiologists. CNN showed higher AUC values than radiologists when diagnosing thyroid nodules (0.66 vs. 0.57, P = 0.04, Table 2). CNN also showed higher values for sensitivity (89.8% vs. 87.3%, P = 0.26), specificity (38.3% vs. 25.5%, P = 0.10), accuracy (83.2% vs. 79.5%, P = 0.08), positive predictive value (90.9% vs. 89.0%, P = 0.07) and negative predictive value (35.3% vs. 22.6%, P = 0.048).

Among 370 nodules, 179 nodules were > 5 mm and 191 nodules were ≤ 5 mm. The characteristics of the patients and nodules are presented in the Supplemental Table S2. Age and portion of malignancy were not different between the subgroups divided by nodule size.

Cut-off values for the malignancy probability from CNN were redefined as > 55.8% for nodules > 5 mm and > 90.3% for nodules ≤ 5 mm. AUC values for diagnosing thyroid nodules did not differ between the CNN and radiologists in nodules > 5 mm (0.69 vs. 0.62, P = 0.25), while CNN showed higher AUC values than radiologists in nodules ≤ 5 mm with borderline significance (0.63 vs. 0.51, P = 0.08, Supplemental Table S3). In nodules ≤ 5 mm, CNN showed lower values for sensitivity (56.8% vs. 92.3%, P < 0.001) and accuracy (58.1% vs. 82.7%, P < 0.001) but higher values for specificity (68.2% vs. 9.1%, P < 0.001).

Comparison of categorization performance between the CNN and radiologists. Among 323 malignant nodules, 4 (1.2%) nodules were category 3, 37 (11.5%) nodules were category 4 and 282 (87.3%) nodules were category 5 according to KSThR TIRADS (Table 1). Among 47 benign nodules, 3 (6.4%) nodules were category 3, 9 (19.2%) nodules were category 4 and 35 (74.5%) nodules were category 5. TIRADS categorization according to CNN showed higher values in the linear trend χ^2-test (20.3 vs. 7.0) and likelihood ratio χ^2-test

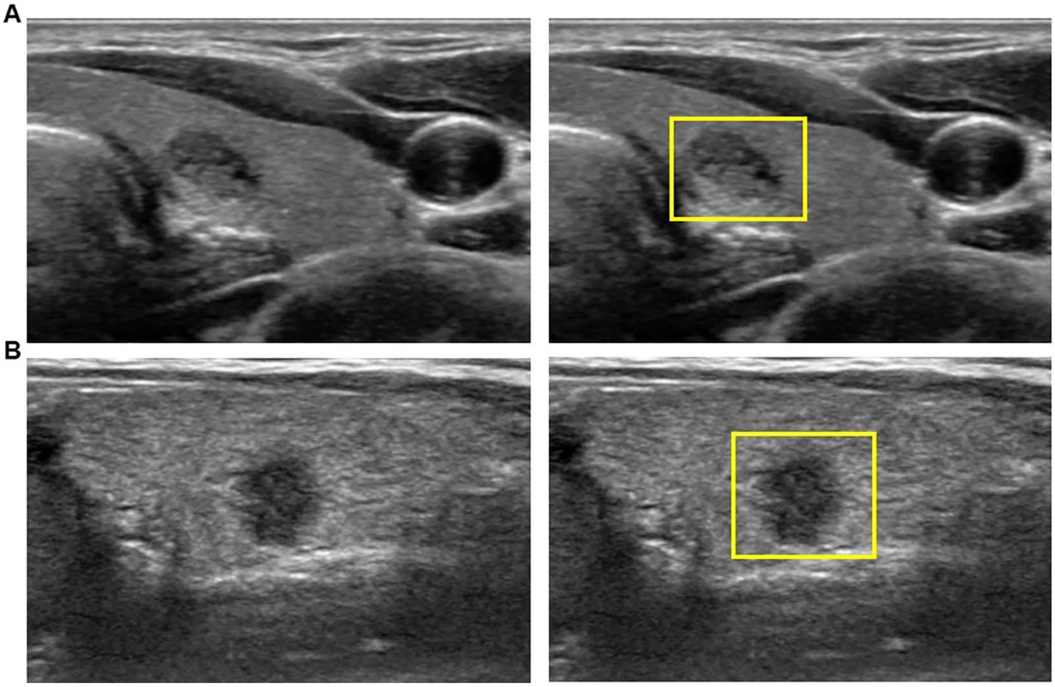

Figure 2. US image of 7 mm-sized thyroid nodules which were later diagnosed as malignant (papillary thyroid microcarcinoma) by surgical histopathology. The square ROI square enclosing the entire targeted thyroid nodule was labeled. (**A**) The nodule was categorized as KSThR TIRADS category 3 due to predominantly solid composition, mild hypoechogenicity, smooth margin, and parallel orientation without microcalcification. The malignancy probability calculated from CNN was 89.3%. (**B**) The nodule was categorized as KSThR TIRADS category 5 due to solid composition, hypoechogenicity, irregular margin, and non-parallel orientation. The malignancy probability calculated from CNN was 96.6%. *US* ultrasound, *KSThR* Korean Society of Thyroid Radiology, *TIRADS* Thyroid Imaging Reporting and Data System, *CNN* convolutional neural network.

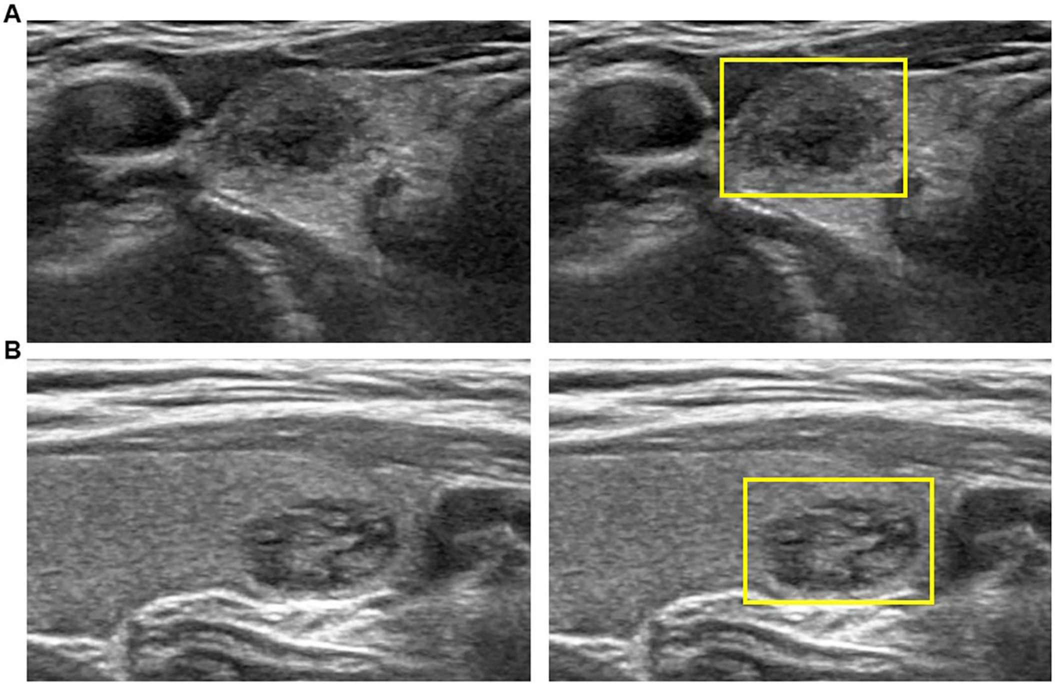

Figure 3. US image of 9 mm-sized thyroid nodules which were later diagnosed as Bethesda category II (benign follicular nodule) by FNA. The square ROI square enclosing the entire targeted thyroid nodule was labeled. (**A**) The nodule was categorized as KSThR TIRADS category 5 due to solid composition, mild hypoechogenicity, and microlobulated margin. The malignancy probability calculated from CNN was 5.8%. (**B**) The nodule was categorized as KSThR TIRADS category 3 due to predominantly solid composition, isoechogenicity, smooth margin, and parallel orientation. The malignancy probability calculated from CNN was 8.8%. *US* ultrasound, *FNA* fine-needle aspiration, *KSThR* Korean Society of Thyroid Radiology, *TIRADS* Thyroid Imaging Reporting and Data System, *CNN* convolutional neural network.

Characteristics	Malignant nodules	Benign nodules	Malignancy rate (%)[d]	*P*-value
No. of patients	317	47		
Age (years)[a]	46.0 ± 12.0	45.9 ± 13.0		0.97
Sex[b]				0.34
Female	251 (79.2%)	40 (85.1%)		
Male	66 (20.8%)	7 (14.9%)		
No. of nodules	323	47		
Nodule size (mm)[c]	5.3 ± 1.5	5.8 ± 2.2		0.14
KSThR TIRADS[c]				0.10
3	4 (1.2%)	3 (6.4%)	57.1 (18.7)	
4	37 (11.5%)	9 (19.2%)	80.4 (5.8)	
5	282 (87.3%)	35 (74.5%)	89 (1.8)	
CNN TIRADS[c]				< 0.001
2	1 (0.3%)	2 (4.3%)	33.3 (27.2)	
3	2 (0.6%)	3 (6.4%)	40 (21.9)	
4	36 (11.1%)	13 (27.7%)	73.5 (6.3)	
5	284 (87.9%)	29 (61.7%)	90.7 (1.6)	

Table 1. Patient demographics and nodal US features. All data except age and malignancy rate are numbers of patients or nodules, with percentages in parentheses. *KSThR* Korean Society of Thyroid Radiology, *TIRADS* Thyroid Imaging Reporting and Data System, *CNN* convolutional neural network. [a]Patient-level comparison using the Student's *t* test for continuous variables. [b]Patient-level comparison using Pearson's χ^2-test for categorical variables. [c]Nodule-level comparison using logistic regression with the generalized estimating equation method. [d]Standard errors are in parentheses.

Performance measures[a]	CNN	Radiologists	P-value
True positive	290	282	
True negative	18	12	
False positive	29	35	
False negative	33	41	
Sensitivity	89.8 (86.5–93.1)	87.3 (83.7–90.9)	0.26
Specificity	38.3 (24.4–52.2)	25.5 (13.1–38)	0.10
Accuracy	83.2 (79.4–87.0)	79.5 (75.3–83.6)	0.08
PPV	90.9 (87.8–94.1)	89.0 (85.5–92.4)	0.07
NPV	35.3 (22.2–48.4)	22.6 (11.4–33.9)	0.048
AUC[b]	0.66 (0.57–0.75)	0.57 (0.50–0.63)	0.04

Table 2. Diagnostic performance of the CNN and radiologists. 95% CIs are noted in parentheses. *CNN* convolutional neural network, *PPV* positive predictive value, *NPV* negative predictive value, *AUC* area under the receiver operating characteristics curve. [a]Each performance measure was compared using logistic regression with the generalized estimating equation method except for AUC. [b]AUC was compared using the Obuchowski algorithm.

(20.9 vs. 6.3) and lower Akaike information criterion values (264.8 vs. 279.4) than KSThR TIRADS assessed by radiologists, suggesting better categorization performance (Table 3).

Discussion

Our study demonstrated that when diagnosing thyroid nodules < 10 mm, CNN trained with thyroid nodules ≥ 10 mm showed better performance than radiologists. CNN also showed better performance than radiologists even in very tiny nodules ≤ 5 mm with borderline significance. In our study, we used a pretrained CNN which was fine-tuned with 13,560 images of thyroid nodules ≥ 10 mm and implemented it to smaller thyroid nodules < 10 mm.

CNN is an end-to-end model that automatically extracts features from digital images to enable pattern recognition, object detection, and classification. Since LeCun et al. proposed LeNet, the first CNN model in 1989, CNN has rapidly developed and various CNNs such as AlexNet or ResNet have been introduced[35]. The CNN-based diagnosis of thyroid nodules has shown comparable performance to experienced radiologists (Table 4). CNN has also shown significantly higher AUC values in recent studies using training sets with large numbers of nodules[19,21,22,25]. In addition, CNN has shown higher specificity than radiologists with similar levels of sensitivity (except in some studies using specific commercially available CAD)[19,21,25].

To the best of our knowledge, no studies have validated the diagnostic performance of CNN on a test set that has a size range different from that of the training set. Our study shows that CNN can diagnose nodules that are completely different in size from those in the training set with significantly better AUC and negative predictive value than experienced radiologists. This is largely consistent with previous studies[19,21]. Our study also shows that differences in specificity and AUC are more significant between the CNN and radiologists in very tiny nodules < 5 mm. Considering the high false-positive rate of FNA in very tiny nodules, we can expect CNN to reduce unnecessary FNA in clinical practice, especially in thyroid micronodules[13].

In our study, the categorization of nodules on CAD values showed comparable or better stratification ability than KSThR TIRADS in terms of discriminatory ability and homogeneity[32–34]. Since the CNN TIRADS defines categories according to the predicted risk of malignancy suggested by KSThR TIRADS, CNN can help clinicians decide the next management step for patients such as whether to follow up or perform FNA under the existing TIRADS guideline. CNN has the potential to be used as a convenient tool that will reduce the burden of clinical triaging thyroid micronodules.

We acknowledge that there are several limitations to our study. First, the number of benign nodules is markedly lower than that of malignant nodules. Because micronodules only underwent FNA when they showed highly suspicious features, FNA-confirmed benign nodules were relatively rare, resulting in low negative predictive value values of both CNN and radiologists. Second, a majority of the malignant nodules were papillary

Test	Linear trend χ^2 test[a]	LR χ^2 test[a]	AIC[b]
CNN TIRADS	20.3	20.9	264.8
KSThR TIRADS	7.0	6.3	279.4

Table 3. Comparison of categorization performance between the CNN and radiologists. *CNN* convolutional neural network, *LR* likelihood ratio, *AIC* Akaike information criterion, *KSThR* Korean Society of Thyroid Radiology, *TIRADS* Thyroid Imaging Reporting and Data System. [a]Higher values suggest better monotonicity of gradient and heterogeneity. [b]Lower values suggest a more parsimonious model.

Author	Training set	Internal test set	External test set	Performances
Wang et al.[25]	5007 nodules	351 nodules, including 151 nodules < 1 cm	N/A	CNN showed significantly higher specificity and AUC than radiologists with comparable sensitivity
				In the subgroup of nodules < 1 cm, CNN also showed significantly higher specificity than radiologists
Li et al.[19]	42,952 patients	1118 patients	1574 patients	CNN showed significantly lower sensitivity and higher specificity than radiologists in both internal and external test sets
Buda et al.[22]	1278 nodules	99 nodules	N/A	CNN showed significantly higher specificity than inexperienced radiologists who did not use ACR TIRADS
				CNN showed similar AUC, sensitivity, and specificity to expert radiologists on the ACR TIRADS committee
Kim et al.[24]	Commercially available CAD	218 nodules ≥ 5 mm	N/A	CNN showed significantly lower specificity and AUC than radiologists with comparable sensitivity
Park et al.[20]	4919 nodules ≥ 5 mm	286 nodules ≥ 5 mm	N/A	No significant difference in diagnostic performance between the CNN and radiologists
Ko et al.[18]	439 nodules ≥ 1 cm and < 2 cm	150 nodules ≥ 1 cm and < 2 cm	N/A	No significant difference in diagnostic performance between the CNN and radiologists
Koh et al.[21]	13,560 nodules ≥ 1 cm	200 nodules ≥ 1 cm	600 nodules ≥ 1 cm	CNN showed significantly higher AUC in the internal test set, while no significant difference was shown in the external test sets
				CNN showed significantly lower sensitivity and higher specificity than radiologists in the internal test set and one of the four external test sets
Han et al.[23]	Commercially available CAD	454 nodules ≥ 1 cm	N/A	CNN showed significantly lower specificity and AUC than radiologists with comparable sensitivity

Table 4. Comparison of diagnostic performance between CNN and radiologists in previous studies. *CNN* convolutional neural network, *N/A* not applicable, *AUC* area under the receiver operating characteristics curve, *ACR* American College of Radiology, *TIRADS* Thyroid Imaging Reporting and Data System, *CAD* computer-aided diagnosis.

thyroid carcinoma. Because follicular neoplasms or the follicular variant of papillary thyroid carcinoma exhibit distinctive US features, our result cannot be generalized to the diagnosis of other pathologic disease entities[36]. Third, radiologists manually selected key images and draw ROIs to be entered into the CNN, implying that the calculations made by CNN are inevitably operator-dependent. In a past study using support vector machine-based CAD, the diagnostic performance of computer-aided diagnosis for thyroid nodules varied significantly according to the experience of radiologists[37,38]. Further studies should be followed to evaluate the reproducibility of CNN.

Conclusion

The deep convolutional neural network trained with thyroid nodules ≥ 10 mm showed overall better diagnostic and categorization performance than radiologists in thyroid nodules < 10 mm, especially those ≤ 5 mm.

Data availability

The raw data analyzed in the study are available from the corresponding author on reasonable request.

Received: 8 September 2022; Accepted: 30 April 2023
Published online: 04 May 2023

References

1. Davies, L., Ouellette, M., Hunter, M. & Welch, H. G. The increasing incidence of small thyroid cancers: Where are the cases coming from?. *Laryngoscope* **120**, 2446–2451. https://doi.org/10.1002/lary.21076 (2010).
2. Mitchell, J. & Parangi, S. The thyroid incidentaloma: An increasingly frequent consequence of radiologic imaging. *Semin. Ultrasound CT MR* **26**, 37–46. https://doi.org/10.1053/j.sult.2004.10.004 (2005).
3. Mandel, S. J. A 64-year-old woman with a thyroid nodule. *JAMA* **292**, 2632–2642. https://doi.org/10.1001/jama.292.21.2632 (2004).
4. Baudin, E. *et al.* Microcarcinoma of the thyroid gland: The Gustave–Roussy Institute experience. *Cancer* **83**, 553–559. https://doi.org/10.1002/(sici)1097-0142(19980801)83:3%3c553::aid-cncr25%3e3.0.co;2-u (1998).
5. Haugen, B. R. *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* **26**, 1–133. https://doi.org/10.1089/thy.2015.0020 (2016).
6. Russ, G. *et al.* European Thyroid Association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: The EU-TIRADS. *Eur. Thyroid J.* **6**, 225–237. https://doi.org/10.1159/000478927 (2017).
7. Shin, J. H. *et al.* Ultrasonography diagnosis and imaging-based management of thyroid nodules: Revised Korean Society of Thyroid Radiology consensus statement and recommendations. *Korean J. Radiol.* **17**, 370–395. https://doi.org/10.3348/kjr.2016.17.3.370 (2016).
8. Tessler, F. N. *et al.* ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White paper of the ACR TI-RADS Committee. *J. Am. Coll. Radiol.* **14**, 587–595. https://doi.org/10.1016/j.jacr.2017.01.046 (2017).
9. Iacobone, M., Jansson, S., Barczynski, M. & Goretzki, P. Multifocal papillary thyroid carcinoma—A consensus report of the European Society of Endocrine Surgeons (ESES). *Langenbecks Arch. Surg.* **399**, 141–154. https://doi.org/10.1007/s00423-013-1145-7 (2014).

10. So, Y. K., Kim, M. W. & Son, Y. I. Multifocality and bilaterality of papillary thyroid microcarcinoma. *Clin. Exp. Otorhinolaryngol.* **8**, 174–178. https://doi.org/10.3342/ceo.2015.8.2.174 (2015).
11. Remonti, L. R., Kramer, C. K., Leitão, C. B., Pinto, L. C. & Gross, J. L. Thyroid ultrasound features and risk of carcinoma: A systematic review and meta-analysis of observational studies. *Thyroid* **25**, 538–550. https://doi.org/10.1089/thy.2014.0353 (2015).
12. Kwak, J. Y., Kim, E. K., Kim, M. J. & Son, E. J. Significance of sonographic characterization for managing subcentimeter thyroid nodules. *Acta Radiol.* **50**, 917–923. https://doi.org/10.1080/02841850903062724 (2009).
13. Mazzaferri, E. L. & Sipos, J. Should all patients with subcentimeter thyroid nodules undergo fine-needle aspiration biopsy and preoperative neck ultrasonography to define the extent of tumor invasion?. *Thyroid* **18**, 597–602. https://doi.org/10.1089/thy.2008.0100 (2008).
14. Moon, H. J., Son, E., Kim, E. K., Yoon, J. H. & Kwak, J. Y. The diagnostic values of ultrasound and ultrasound-guided fine needle aspiration in subcentimeter-sized thyroid nodules. *Ann. Surg. Oncol.* **19**, 52–59. https://doi.org/10.1245/s10434-011-1813-1 (2012).
15. Shin, H. C. *et al.* Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* **35**, 1285–1298. https://doi.org/10.1109/tmi.2016.2528162 (2016).
16. Tajbakhsh, N. *et al.* Convolutional neural networks for medical image analysis: Full training or fine tuning?. *IEEE Trans. Med. Imaging* **35**, 1299–1312. https://doi.org/10.1109/tmi.2016.2535302 (2016).
17. Ma, J., Wu, F., Zhu, J., Xu, D. & Kong, D. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. *Ultrasonics* **73**, 221–230. https://doi.org/10.1016/j.ultras.2016.09.011 (2017).
18. Ko, S. Y. *et al.* Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound. *Head Neck* **41**, 885–891. https://doi.org/10.1002/hed.25415 (2019).
19. Li, X. *et al.* Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: A retrospective, multicohort, diagnostic study. *Lancet Oncol.* **20**, 193–201. https://doi.org/10.1016/s1470-2045(18)30762-9 (2019).
20. Park, V. Y. *et al.* Diagnosis of thyroid nodules: Performance of a deep learning convolutional neural network model vs. radiologists. *Sci. Rep.* **9**, 17843. https://doi.org/10.1038/s41598-019-54434-1 (2019).
21. Koh, J. *et al.* Diagnosis of thyroid nodules on ultrasonography by a deep convolutional neural network. *Sci. Rep.* **10**, 15245. https://doi.org/10.1038/s41598-020-72270-6 (2020).
22. Buda, M. *et al.* Management of thyroid nodules seen on US images: Deep learning may match performance of radiologists. *Radiology* **292**, 695–701. https://doi.org/10.1148/radiol.2019181343 (2019).
23. Han, M., Ha, E. J. & Park, J. H. Computer-aided diagnostic system for thyroid nodules on ultrasonography: Diagnostic performance based on the thyroid imaging reporting and data system classification and dichotomous outcomes. *Am. J. Neuroradiol.* https://doi.org/10.3174/ajnr.A6922 (2020).
24. Kim, H. L., Ha, E. J. & Han, M. Real-world performance of computer-aided diagnosis system for thyroid nodules using ultrasonography. *Ultrasound Med. Biol.* **45**, 2672–2678. https://doi.org/10.1016/j.ultrasmedbio.2019.05.032 (2019).
25. Wang, L. *et al.* Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network. *World J. Surg. Oncol.* **17**, 12. https://doi.org/10.1186/s12957-019-1558-z (2019).
26. Kim, E. K. *et al.* New sonographic criteria for recommending fine-needle aspiration biopsy of nonpalpable solid nodules of the thyroid. *AJR Am. J. Roentgenol.* **178**, 687–691. https://doi.org/10.2214/ajr.178.3.1780687 (2002).
27. Kwak, J. Y. *et al.* Thyroid imaging reporting and data system for US features of nodules: A step in establishing better stratification of cancer risk. *Radiology* **260**, 892–899. https://doi.org/10.1148/radiol.11110206 (2011).
28. Yoon, J. H., Lee, H. S., Kim, E. K., Moon, H. J. & Kwak, J. Y. Malignancy risk stratification of thyroid nodules: Comparison between the thyroid imaging reporting and data system and the 2014 American Thyroid Association management guidelines. *Radiology* **278**, 917–924. https://doi.org/10.1148/radiol.2015150056 (2016).
29. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), *Las Vegas,* 770–778. https://doi.org/10.1109/CVPR.2016.90 (2016).
30. He, K., Zhang, X., Ren, S. & Sun, J. Identity Mappings in Deep Residual Networks ArXiv Web site. arXiv:1603.05027 (2016).
31. Obuchowski, N. A. Nonparametric analysis of clustered ROC curve data. *Biometrics* **53**, 567–578 (1997).
32. An, C., Choi, G. H., Lee, H. S. & Kim, M. J. Assessment of preoperative magnetic resonance imaging staging in patients with hepatocellular carcinoma undergoing resection compared with the seventh American Joint Committee on Cancer System. *Investig. Radiol.* **47**, 634–641. https://doi.org/10.1097/RLI.0b013e3182630e8d (2012).
33. Marrero, J. A. *et al.* Prognosis of hepatocellular carcinoma: Comparison of 7 staging systems in an American cohort. *Hepatology* **41**, 707–716. https://doi.org/10.1002/hep.20636 (2005).
34. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**, 716–723. https://doi.org/10.1109/TAC.1974.1100705 (1974).
35. LeCun, Y. *et al.* Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1**, 541–551. https://doi.org/10.1162/neco.1989.1.4.541 (1989).
36. Yoon, J. H., Kwon, H. J., Kim, E. K., Moon, H. J. & Kwak, J. Y. The follicular variant of papillary thyroid carcinoma: Characteristics of preoperative ultrasonography and cytology. *Ultrasonography* **35**, 47–54. https://doi.org/10.14366/usg.15037 (2016).
37. Jeong, E. Y. *et al.* Computer-aided diagnosis system for thyroid nodules on ultrasonography: Diagnostic performance and reproducibility based on the experience level of operators. *Eur. Radiol.* **29**, 1978–1985. https://doi.org/10.1007/s00330-018-5772-9 (2019).
38. Zhou, Y. *et al.* Differentiating non-lactating mastitis and malignant breast tumors by deep-learning based AI automatic classification system: A preliminary study. *Front. Oncol.* **12**, 997306. https://doi.org/10.3389/fonc.2022.997306 (2022).

Acknowledgements

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (2021R1A2C2007492).

Author contributions

M.R., S.H.C. and J.Y.K. contributed to the study design and writing the first draft of manuscript. E.L. developed the AI-CAD program and collected data from the program. Statistical analysis of data was performed by H.S.L. and K.H. All authors reviewed, contributed to, and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-34459-3.

Correspondence and requests for materials should be addressed to J.Y.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.