Diagnosis of thyroid micronodule on ultrasound using deep convolutional neural network

Chun, Sei Hyun

Department of Medicine

The Graduate School, Yonsei University

Diagnosis of thyroid micronodule on ultrasound using deep convolutional neural network

Chun, Sei Hyun

Department of Medicine

The Graduate School, Yonsei University

Diagnosis of thyroid micronodule on ultrasound using deep convolutional neural network

Directed by Professor Kwak, Jin Young

The Master's Thesis
submitted to the Department of Medicine,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree of
Master of Medical Science

Chun, Sei Hyun

December 2020

This certifies that the Master's Thesis of
Chun, Sei Hyun is approved.

Thesis Supervisor : Kwak, Jin Young

Thesis Committee Member#1 : Yoon, Jung Hyun

Thesis Committee Member#2 : Han, Kyunghwa

The Graduate School
Yonsei University

December 2020

ACKNOWLEDGEMENTS

First, I would like to appreciate Professor Jin Young Kwak for generous support and academic guidance throughout the whole research. Also, I would like to appreciate Professor Jung Hyun Yoon and Professor Kyunghwa Han for their valuable advice on the thesis. Sincere thanks to my family who has always been the greatest strength.

<TABLE OF CONTENTS>

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Diagnosis of thyroid micronodule on ultrasound using deep convolutional neural network

Chun, Sei Hyun

Department of Medicine
The Graduate School, Yonsei University

(Directed by Professor Kwak, Jin Young)

Backgrounds: We implemented computer-aided diagnosis based on a convolutional neural network (CNN) model to investigate the performance of CNN to discriminate malignant from benign thyroid nodules measuring < 10 mm. We also compared the diagnostic performance of CNN with those of radiologists.

Methods: The CNN was trained by using ultrasound (US) images of 13,560 nodules measuring ≥ 10 mm collected from Severance Hospital. Between March 2016 and February 2018, US images of 370 nodules measuring < 10 mm from 362 consecutive patients were retrospectively collected from Severance Hospital. All nodules were confirmed as malignant or benign from aspirate cytology or surgical histology. The diagnostic performance of CNN and radiologists were assessed and compared in area under curve (AUC), sensitivity, specificity, accuracy, positive predictive value, and negative predictive value (NPV). Subgroup analysis were performed based on the nodule size with cutoff value of 5 mm. The categorization performance of CNN and radiologists were also compared.

Results: Among 370 nodules, 323 nodules were malignant, and 47 nodules were benign. CNN showed significantly higher NPV (35.3 vs. 22.6, $P=0.048$) and AUC (0.663 vs. 0.567, $P = 0.045$) than radiologists. CNN also showed better categorization performance than radiologists. In subgroup of nodules measuring $\leq$ 5 mm, CNN showed higher AUC (0.629 vs. 0.507, $P=0.077$) and specificity (68.2% vs. 9.1%, $P<0.001$) than radiologists.

Conclusion: CNN trained with thyroid nodules ≥ 10 mm showed overall better diagnostic performance with radiologists in diagnosis and categorization of thyroid nodules < 10 mm, especially in nodules $\leq$ 5mm.

Key words : artificial intelligence, thyroid nodule, ultrasonography

Diagnosis of thyroid micronodule on ultrasound using deep convolutional neural network

Chun, Sei Hyun

Department of Medicine
The Graduate School, Yonsei University

(Directed by Professor Kwak, Jin Young)

I. INTRODUCTION

Detection of thyroid nodule has substantially increased with widespread use of high-resolution ultrasound (US), resulting in high prevalence of 19–67% in general population[1-4]. Among them, approximately 7–15% of them are found to be thyroid cancers[5,6]. The recommendation of fine-needle aspiration (FNA) have been controversial in thyroid micronodules (< 10 mm) because most patients with papillary thyroid microcarcinoma (PTMC), defined as tumor < 10 mm, had near-zero cancer-specific mortality[7]. Therefore, many guidelines suggested active surveillance for thyroid nodules < 10 mm as well as FNA as an available option, depending on clinical settings and patient preferences[8-12].

Multifocality and bilaterality in papillary thyroid carcinoma are common features with reported frequency as 18–87%[13], and are known to be risk factors of nodal metastasis, distant metastasis and regional recurrence after initial therapy[14]. The American Thyroid Association guideline recommended that lobectomy should be initially performed for unifocal PTMC without extrathyroidal extension, but also noted that the presence of bilateral nodule can be recommending criteria for bilateral thyroidectomy to address the possibility of bilaterality[9]. Considering that physician’s visual analysis at micronodules, especially nodules smaller than 5 mm, have been reported to show high false positive rate of US, preoperative detection of very tiny nodules may increase additional FNA[15,16]. Given the high nondiagnostic rate of FNA in very tiny nodules, preoperative diagnosis is a challenging task[14,17].

Convolutional neural network (CNN) is a type of deep learning models which enables high-performance visual recognition and classification after automatically learning representative features from training set[18,19]. The characteristics of training set is therefore critical to the performance of CNN. To differentiate malignant thyroid nodules from benign nodules efficiently, CNN-based method has been investigated[20-25]. Some investigations showed the results of validation for nodules corresponding to the same size criteria with the training set[21,24,25], while no other studies demonstrate the nodule size for training or validation of CNN[20,22,23]. To the best of our knowledge, no studies has applied a CNN-based model to nodules beyond the size criteria of training set. In this study, we investigated diagnostic performances of CNN which were trained with nodules ≥ 10 mm at thyroid nodules < 10 mm and compared with those of radiologists.

II. MATERIALS AND METHODS

The institutional review board of our institution approved this retrospective study, with a waiver for informed consent. Signed informed consent of biopsy or surgical procedures was obtained preoperatively from all patients.

1. Patients

This study was performed at Severance Hospital (a tertiary referral center) from March 2016 to February 2018, during which US-guided FNA was consulted for 4110 nodules in 3716 consecutive patients. Initial FNA was performed in 3323 nodules in 3240 patients, of which 698 nodules were < 10 mm in 683 patients. Out study included nodules < 10 mm if they (a) were cytologically confirmed benign or malignant (Bethesda category II or VI) or (b) were confirmed from postsurgical histologies. Finally a total of 370 nodules in 362 patients (mean ages, 46.1 years ± 12.2; range 20–76 years), including 289 (79.8%; mean ages, 46.4 years ± 12.3; range, 20–76 years) women and 73 (20.2%; mean ages, 45.0 years ± 11.9; range, 26–73 years) men were included (Fig. 1). There were 347 (93.8%) nodules which were confirmed with surgery and 23 (6.2%) nodules which were confirmed with FNA. The reasons for FNA in 370 nodules enrolled were phisicians‘ requests from outside clinics (n=127), high suspicion nodule > 5 mm (n=123)[11], determination of the surgical extent in patients with bilateral nodules (n=83), patient requests (n=30) and cervical lymph node metastasis (n=7).

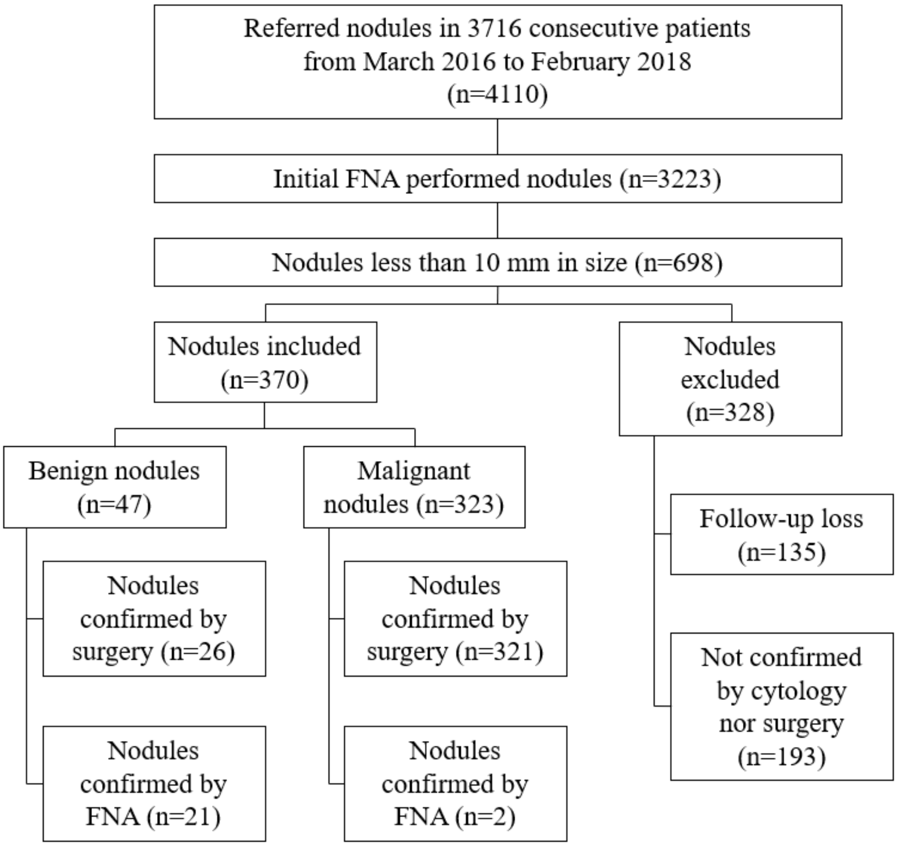
Referred nodules in 3716 consecutive patients from March 2016 to February 2018 (n=4110)
Initial FNA performed nodules (n=3223)
Nodules less than 10 mm in size (n=698)
Nodules included (n=370)
Nodules excluded (n=328)
Benign nodules (n=47)
Malignant nodules (n=323)
Nodules confirmed by surgery (n=26)
Nodules confirmed by FNA (n=21)
Nodules confirmed by surgery (n=321)
Nodules confirmed by FNA (n=2)
Follow-up loss (n=135)
Not confirmed by cytology nor surgery (n=193)

Figure 1. Flow chart of patient enrollment. Total 370 nodules including 323 malignant nodules and 47 benign nodules were included in this study. FNA: fine needle aspiration.

2. 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 thyroid nodules and recorded US features prospectively in our institutional database[26,27]. The US features including composition, echogenicity, margin, calcifications, and shape were recorded using descriptors which have been used from June 2012 to the present in our institution[28]. An experienced radiologist (K.J.Y with 20 years of experience dedicated in thyroid imaging) who was blinded to clinical information and pathological results reassigned Korean Society of Thyroid Radiology (KSThR) Thyroid Imaging Reporting and Data System (TIRADS) categories to each thyroid nodule according to the pre-recorded US features[11].

3. Image Acquisition and CNN evaluation

An experienced radiologist (K.J.Y) selected and retrieved a representative US image for each thyroid nodule from the picture archiving and communication system and stored it as JPEG formats. For each image, a square region of interest (ROI) enclosing the entire targeted thyroid nodule was manually labeled using the Paint program of Window 10 by the radiologist (K.J.Y). The ROIs were extracted to calculate the percent of malignancy by the CNN per each thyroid nodule[21].

We developed a computer-aided diagnosis (CAD) program to assess the risk of malignancy of thyroid nodules on US images. The CAD program, using pretrained CNN model ResNetV2, was trained with 13,560 US images of thyroid nodules which were cytologically or surgically proven to be either malignant or benign[25]. All nodules in the training set were measuring 10mm or larger in size, consisted of 7,160 malignancy and 6,400 benign nodules. Using the CAD program, we calculated the risks of malignancy as continuous values ranging from 0-100% (CAD value).

We also categorized nodules by designating category based on the CAD value (CNN TIRADS) according to the predicted probability from KSThR TIRADS (Figures 2 and 3). CNN TIRADS category 2 was designated for nodules with malignancy probability < 3%, category 3 for probability < 15%, category 4 for probability < 60% and category 5 for probability ≥ 60%[11].

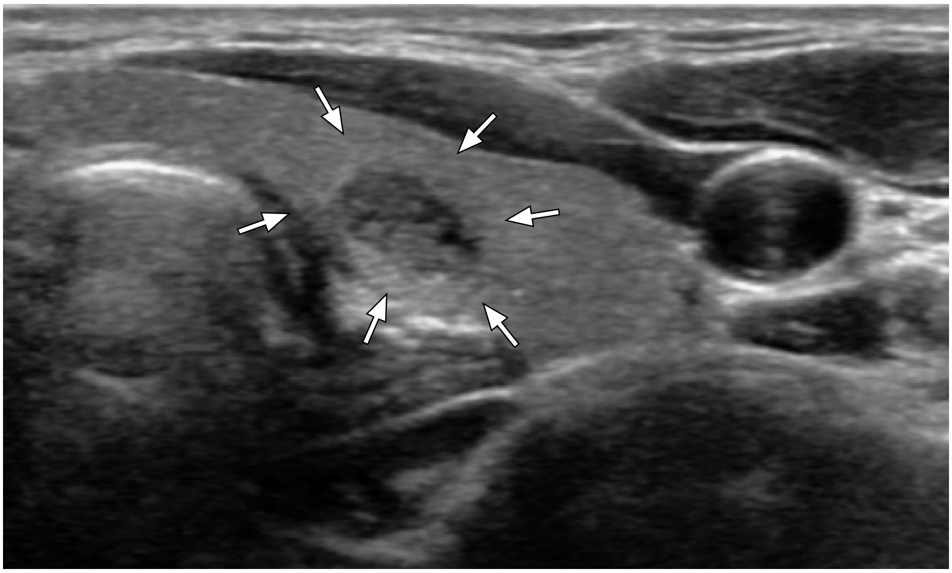

Figure 2. US image of an about 7mm-sized thyroid nodule (white arrows) which was later diagnosed as malignant (papillary thyroid microcarcinoma) by surgical histopathology. This nodule was categorized as KSThR TIRADS category 3 due to predominantly solid composition, mild hypoechogenicity, smooth margin and parallel orientation without microcalcification. The malignancy probability calculated from CNN was 89.3%. US: ultrasound, KSThR: Korean Society of

Thyroid Radiology, TIRADS: Thyroid Imaging Reporting and Data System, CNN: convolutional neural network.

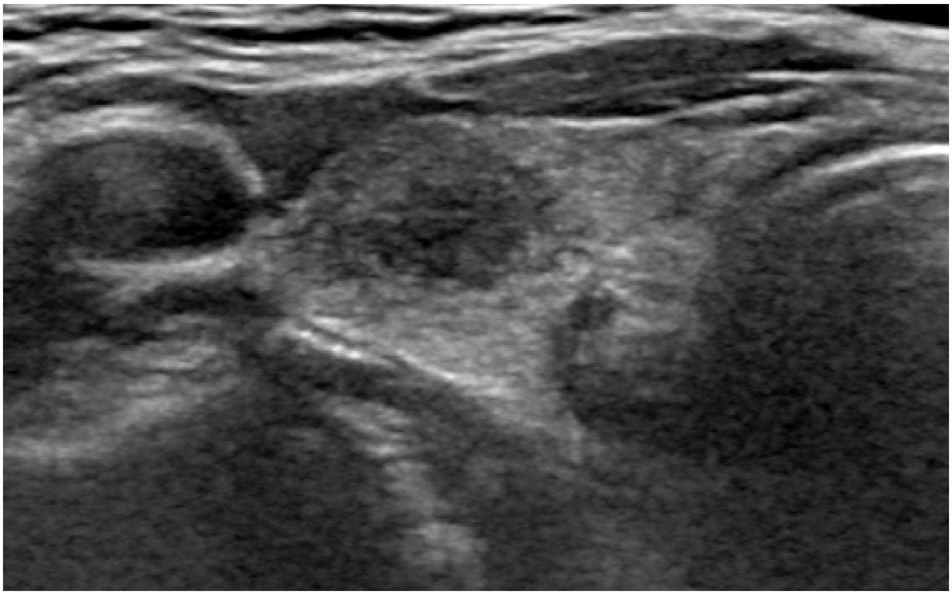

Figure 3. US image of an about 9mm-sized thyroid nodule (white arrows) which was later diagnosed as Bethesda category II. (benign follicular nodule) by FNA. This nodule was categorized as KSThR TIRADS category 5 due to solid composition, mild hypoechogenicity and microlobulated margin. The malignancy probability calculated from CNN was 5.8%. US: ultrasound, FNA: fine needle aspiration, KSThR: Korean Society of Thyroid Radiology, TIRADS: Thyroid Imaging Reporting and Data System, CNN: convolutional neural network.

4. Statistical analysis

For the reference standard, histopathologic result from FNA or surgery was used to confirm the final diagnosis of each thyroid nodule. If a nodule underwent both FNA and surgery or there was a discrepancy between the two tests, the reference standard was set to the histopathologic result from surgical specimen.

Baseline characteristics of patients and US features of nodules were compared between malignant nodules and benign nodules by using the Student's *t*-test and Pearson's χ^2-test in patient-level and logistic regression analysis with generalized estimating equations (GEE) method for clustered data in nodule-level comparison. For statistical analysis, receiver operating characteristic curve analysis was used to obtain area under curve (AUC) with 95% confidence intervals and the TIRADS category and CAD value evaluated on each thyroid nodule were divided into positive and negative according to the Youden index. We assessed and compared the diagnostic performances of the TIRADS category and CNN including sensitivity, specificity, accuracy, positive predictive value (PPV) and negative predictive value (NPV) by using logistic regression with GEE method. The AUC values were compared between the CNN and radiologists using the DeLong algorithm[29]. Subgroup analysis was performed separately according to the nodule size with cutoff value > 5mm. Same statistics analysis was performed for subgroup analysis.

We assessed categorization performance of CNN TIRADS and KSThR TIRADS by using linear trend χ^2-test and the likelihood ratio (LR) χ^2-test to determine discriminatory ability (small differences in risk of malignancy among nodules in the same category), homogeneity (greater differences in risk of malignancy among nodules in the different category) and monotonicity of gradients (whether the risk of malignancy of nodules increases as the category increases) of each categorization system[30,31]. We also used Akaike information criterion (AIC), which is a widely used estimator for model selection. Smaller AIC values indicates the more informative model[32].

Statistical analysis was performed using statistical software (SAS version 9.4, SAS Institute, Cary, NC, USA). Two-sided *P* values <0.05 were considered to indicate statistical significance.

III. RESULTS

1. Patients and nodules characteristics

Among the 370 enrolled nodules, 323 nodules were confirmed malignant and 47 nodules were confirmed benign. Of these malignant nodules, 322 nodules were confirmed as papillary thyroid carcinoma and 1 nodule was confirmed as medullary thyroid carcinoma. The mean nodule size of malignant and benign nodules were 5.3 ± 1.5 mm and 5.8 ± 2.2 mm, respectively (P=0.144, Table 1). No significant difference was observed between the malignant nodules and the benign nodules in age (46.0 years vs. 45.9 years, P=0.971) and proportion of female (79.2% vs. 85.1%, P=0.344).

Table 1. Demographics of patients and US features of nodules

Characteristics	Malignant nodules	Benign nodules	Malignancy rate (SE) (%)	*P*-value
No. of patients	317	47		
Age(years)[a]	46.0 ± 12.0	45.9 ± 13.0		0.971
Sex[b]				0.344
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		0.144
KSThR TIRADS[c]				0.096
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)	

All data except age, nodule size and malignancy rate are numbers of patients or nodules, with the percentage in parentheses.

Age and nodule size are expressed as the means ± SEs.

[a] patient-level comparison by using Student's *t*-test for continuous variable.

[b] patient-level comparison by using Pearson's χ^2-test for categorical variable.

[c] nodule-level comparison by using logistic regression with GEE method.

US: ultrasound, SE: standard error, KSThR: Korean Society of Thyroid Radiology, TIRADS: Thyroid imaging reporting and data system, CNN: convolutional neural network, GEE: generalized estimating equations, CNN TIRADS: Categorization of CAD values according to predicted value per category from KSThR TIRADS.

2. Comparison of diagnostic performance

The optimal cut-off points, set by Youden index, were probability >56.1 for CNN and KSThR TIRADS category 5 for radiologists. CNN showed significantly higher AUC value than the radiologists to diagnose thyroid nodules (0.663 vs. 0.567, P = 0.045, Table 2). CNN showed higher values of sensitivity (89.8% vs. 87.3%, P = 0.257), specificity (38.3% vs. 25.5%, P = 0.099), accuracy (83.2% vs. 79.5%, P = 0.079), PPV (90.9% vs. 89.0%, P = 0.072) and NPV (35.3% vs. 22.6%, P = 0.048).

Table 2. Comparison of diagnostic performance

Performance measures[a]	CNN	Radiologists	*P*-value
TP	290	282	
TN	18	12	
FP	29	35	
FN	33	41	
Sensitivity	89.8 (86.5-93.1)	87.3 (83.7-90.9)	0.257
Specificity	38.3 (24.4-52.2)	25.5 (13.1-38)	0.099
Accuracy	83.2 (79.4-87.0)	79.5 (75.3-83.6)	0.079
PPV	90.9 (87.8-94.1)	89.0 (85.5-92.4)	0.072
NPV	35.3 (22.2-48.4)	22.6 (11.4-33.9)	0.048
AUC[b]	0.663 (0.571-0.754)	0.567 (0.5-0.633)	0.045

95% confidence intervals are noted in parentheses.

[a] Each performance measure was compared by using logistic regression with GEE method except AUC.

[b] AUC was compared by using DeLong algorithm.

CNN: convolutional neural network, PPV: positive predictive value, NPV: negative predictive value, AUC: area under curve, GEE: generalized estimating equations.

Among 370 nodules, 179 nodules had size > 5mm and 191 nodules had size ≤ 5mm. The patients and nodules characteristics were demonstrated in the Table 3. The age and portion of malignancy were not significantly different between the subgroups based on nodule size.

Cutoff values for malignancy probability from CNN were redefined as > 55.8% for nodules > 5 mm and > 90.3% for nodules ≤ 5 mm. The AUC to diagnose thyroid nodules showed no significant difference between CNN and radiologists in nodules > 5 mm (0.693 vs. 0.615, P = 0.251), while CNN showed higher AUC than radiologists in nodules ≤ 5 mm with borderline significance (0.629 vs. 0.507, P=0.077, Table 4). In nodules ≤ 5 mm, CNN showed significantly lower values of sensitivity (56.8% vs. 92.3%, P < 0.001) and accuracy (58.1% vs. 82.7%, P < 0.001) but significantly higher values of specificity (68.2% vs. 9.1%, P < 0.001).

Table 3. Demograhics of patients and US features of nodules in subgroup

Characteristics	Nodules > 5mm	Nodules ≤ 5mm	*P*-value
No. of patients	177	188	
Age (years)[a]	47.5 ± 12.2	44.7 ± 11.9	0.027
Sex[b]			0.53
Female	144 (81.4%)	148 (78.7%)	
Male	33 (18.6%)	40 (21.3%)	
No. of nodules	179	191	
Nodule size (mm)[c]	6.7 ± 1.0	4.1 ± 0.9	<.001

Nodular pathology [c]			0.477
Malignant	154 (86%)	169 (88.5%)	
Benign	25 (14%)	22 (11.5%)	
KSThR TIRADS[c]			<.001
3	7 (3.9%)	0 (0%)	
4	31 (17.3%)	15 (7.9%)	
5	141 (78.8%)	176 (92.2%)	
CNN TIRADS[c]			
2	2 (1.1%)	1 (0.5%)	
3	5 (2.8%)	0 (0%)	
4	31 (17.3%)	18 (9.4%)	
5	141 (78.8%)	172 (90.1%)	

All data except age and nodule size are numbers of patients or nodules, with the percentage in parentheses.

Age and nodule size are expressed as the means ± SEs.

[a] patient-level comparison by using Student's *t*-test for continuous variable.

[b] patient-level comparison by using Pearson's χ^2-test for categorical variable.

[c] nodule-level comparison by using logistic regression with GEE method.

US: ultrasound, KSThR: Korean Society of Thyroid Radiology, TIRADS: Thyroid imaging reporting and data system, CNN: convolutional neural network, GEE: generalized estimating equations, CNN TIRADS: Categorization of CAD values according to predicted value per category from KSThR TIRADS, SE: standard error.

Table 4. Comparison of diagnostic performance in subgroup

Performance measures[a]	CNN	Radiologists	p-value
	Nodules measuring > 5mm (n=179)		
TP	131	126	
TN	13	10	
FP	12	15	
FN	23	28	
Sensitivity	85.1 (79.4-90.7)	81.8 (75.7-87.9)	0.369
Specificity	52.0 (32.4-71.6)	40.0 (20.8-59.2)	0.307
Accuracy	80.4 (74.6-86.3)	76.0 (69.7-82.2)	0.204
PPV	91.6 (87.1-96.2)	89.4 (84.3-94.5)	0.250
NPV	36.1 (20.4-51.8)	26.3 (12.3-40.3)	0.179
AUC[b]	0.693 (0.566-0.819)	0.615 (0.509-0.72)	0.251
	Nodules measuring ≤ 5mm (n=191)		
TP	159	156	
TN	5	2	
FP	17	20	

FN	10	13	
Sensitivity	56.8 (49.3-64.3)	92.3 (88.3-96.3)	<.0001
Specificity	68.2 (48.7-87.6)	9.1 (0-21.1)	<.0001
Accuracy	58.1 (51.1-65.1)	82.7 (77.4-88.1)	<.0001
PPV	93.2 (88.3-98.1)	88.6 (83.9-93.3)	0.036
NPV	17.0 (9.2-24.9)	13.3 (0-30.5)	0.652
AUC[b]	0.629 (0.497-0.761)	0.507 (0.442-0.572)	0.077

95% confidence intervals are noted in parentheses.

[a] Each performance measure was compared by using logistic regression with GEE method except AUC.

[b] AUC was compared by using DeLong algorithm.

CNN: convolutional neural network, PPV: positive predictive value, NPV: negative predictive value, AUC: area under curve, GEE: generalized estimating equations.

3. Comparison of categorization performance

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. 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 linear trend χ^2-test (20.3 vs. 7.0) and LR χ^2-test (20.9 vs. 6.3) and lower AIC values (264.8 vs.

279.4) than KSThR TIRADS assessed by radiologists, suggesting better categorization performance (Table 5).

Table 5. Comparison of categorization performance

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

[a] Higher values suggest better discriminatory ability and homogeneity.

[b] Lower values suggest preferred model.

CNN: convolutional neural network, LR: likelihood ratio, AIC: Akaike information criterion, TIRADS: Thyroid imaging reporting and data system, KSThR: Korean Society of Thyroid Radiology, CNN TIRADS: Categorization of CAD values according to predicted value per category from KSThR TIRADS.

IV. DISCUSSION

Our study demonstrated that in diagnosis of 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 CNN which was pretrained with 1,281,167 non-medical images and fine-tuned with 13,560 images of thyroid nodules ≥ 10 mm[25].

CNN is an end-to-end model that automatically extract 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 been rapidly developing and various CNNs such as AlexNet or ResNet have been developed[33]. To classify an image through CNN, feature maps are extracted via convolution layers, spatial dimensions are reduced via pooling layers, and fully connected multilayer perceptron finally provides probability for each class. CNN-based diagnosis of thyroid nodules has been reported to show comparable performance to experienced radiologists (Table 6). CNN have showed significantly higher AUC in some recent studies using training sets with large numbers of nodules[22,25,34]. CNN also have shown higher specificity than radiologists with similar level of sensitivity (except some studies using specific commercially available CAD)[22,23,25,35,36].

Table 6. Comparison of diagnostic performance between CNN and radiologists in previous studies

Author	Training set	Internal test set	External test set	Performances
Wang L et al.[22]	5007 nodules	351 nodules, including 151 nodules < 1 cm	N/A	CNN showed significantly higher specificity and AUC than radiologists with comparable sensitivity. In subgroup of nodules < 1 cm, CNN also showed significantly higher specificity than radiologists.
Li X et al.[23]	42952 patients	1118 patients	1574 patients	CNN showed significantly lower sensitivity and higher specificity than radiologists in both internal and external test sets.

Buda M et al.[34]	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 in ACR TIRADS committee.
Kim HL et al.[35]	Commercially available CAD	218 nodules ≥ 5mm	N/A	CNN showed significantly lower specificity and AUC than radiologists with comparable sensitivity.
Park VY et al.[24]	4919 nodules ≥ 5 mm	286 nodules ≥ 5 mm	N/A	No significant difference in diagnostic performance between CNN and radiologists.
Ko SY et al.[21]	439 nodules ≥ 1 cm and < 2 cm	150 nodules ≥ 1 cm and < 2 cm	N/A	No significant difference in diagnostic performance between CNN and radiologists.
Koh J et al.[25]	13560 nodules ≥ 1 cm	200 nodules ≥ 1 cm	600 nodules ≥ 1 cm	CNN showed significantly higher AUC in internal test set, while no significant difference was shown in external test sets. CNN showed significantly lower sensitivity and higher specificity than radiologists in both internal and one of the four external test sets.
Han M et al.[36]	Commercially available CAD	454 nodules ≥ 1 cm	N/A	CNN showed significantly lower specificity and AUC than radiologists with comparable sensitivity.

CNN: convolutional neural network, N/A: not applicable, AUC: area under curve, ACR: American College of Radiology, TIRADS: Thyroid imaging reporting and data system, CAD: computer-aided diagnosis.

To the best of our knowledge, no studies have validated the diagnostic performance of CNN on test set of a size range different from that of the training set. Our study shows that CNN can diagnose nodules which is completely different from those in training set in aspect of size with significantly better AUC and NPV than experienced radiologists. This is largely consistent with previous studies[22,23,25]. Our study also shows that the differences of specificity and AUC between CNN and radiologists are more significant in very tiny nodules < 5 mm. Considering high false positive rate of FNA in very tiny nodules, using CNN can reduce unnecessary FNA in clinical practice, especially in thyroid micronodules[15].

In our study, categorization of nodules on CAD values show comparable or better stratification ability than KSThR TIRADS in aspect of discriminatory ability and homogeneity[30-32]. Since the CNN TIRADS is defined according to the predicted risk of malignancy per category from KSThR TIRADS, CNN can help to decide the next step such as whether to follow up or perform FNA under the existing TIRADS guideline. CNN may be a convenient tool for radiologists to reduce the burden for clinical triaging of thyroid micronodules.

We acknowledge that there are several limitations in our study. First, the number of benign nodules is markedly lower than that of malignant nodules. Because FNA was performed to micronodules only when they showed highly suspicious features, FNA-confirmed benign nodules were relatively rare. Second, majority of malignant nodules were papillary thyroid carcinoma. Because follicular neoplasms or follicular variant of papillary thyroid carcinoma exhibit distinctive US features, our result is not generalized to the diagnosis of other pathologic disease entities[37]. Third, radiologists manually select the key image and draw the ROI to be entered into the CNN, suggesting that the calculation of CNN inevitably implies operator dependency. Diagnostic performance of computer-aided diagnosis of thyroid nodule has been reported to vary significantly, depending on the experience of radiologists in a study using support vector machine-based CAD[38]. Further study should be followed to evaluate the reproducibility of CNN.

V. CONCLUSION

CNN trained with thyroid nodules ≥ 10 mm showed overall better diagnostic and categorization performance than radiologists in thyroid nodules < 10 mm, especially in nodules ≤ 5mm.

REFERENCES

1. Davies L, Ouellette M, Hunter M, Welch HG. The increasing incidence of small thyroid cancers: where are the cases coming from? Laryngoscope 2010;120:2446-51.

2. Guth S, Theune U, Aberle J, Galach A, Bamberger CM. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. Eur J Clin Invest 2009;39:699-706.

3. Mitchell J, Parangi S. The thyroid incidentaloma: an increasingly frequent consequence of radiologic imaging. Semin Ultrasound CT MR 2005;26:37-46.

4. Tan GH, Gharib H. Thyroid incidentalomas: management approaches to nonpalpable nodules discovered incidentally on thyroid imaging. Ann Intern Med 1997;126:226-31.

5. Hegedus L. Clinical practice. The thyroid nodule. N Engl J Med 2004;351:1764-71.

6. Mandel SJ. A 64-year-old woman with a thyroid nodule. Jama 2004;292:2632-42.

7. Baudin E, Travagli JP, Ropers J, Mancusi F, Bruno-Bossio G, Caillou B, et al. Microcarcinoma of the thyroid gland: the Gustave-Roussy Institute experience. Cancer 1998;83:553-9.

8. Gharib H, Papini E, Garber JR, Duick DS, Harrell RM, Hegedus L, et al. AMERICAN ASSOCIATION OF CLINICAL ENDOCRINOLOGISTS, AMERICAN COLLEGE OF ENDOCRINOLOGY, AND ASSOCIAZIONE MEDICI ENDOCRINOLOGI MEDICAL

GUIDELINES FOR CLINICAL PRACTICE FOR THE DIAGNOSIS AND MANAGEMENT OF THYROID NODULES--2016 UPDATE. Endocr Pract 2016;22:622-39.

9. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, 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.

10. Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R, Leenhardt L. European Thyroid Association Guidelines for Ultrasound Malignancy Risk Stratification of Thyroid Nodules in Adults: The EU-TIRADS. Eur Thyroid J 2017;6:225-37.

11. Shin JH, Baek JH, Chung J, Ha EJ, Kim JH, Lee YH, et al. Ultrasonography Diagnosis and Imaging-Based Management of Thyroid Nodules: Revised Korean Society of Thyroid Radiology Consensus Statement and Recommendations. Korean J Radiol 2016;17:370-95.

12. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, 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.

13. 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 2014;399:141-54.

14. So YK, Kim MW, Son YI. Multifocality and bilaterality of papillary thyroid microcarcinoma. Clin Exp Otorhinolaryngol 2015;8:174-8.

15. Mazzaferri EL, 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 2008;18:597-602.

16. Moon HJ, Son E, Kim EK, Yoon JH, Kwak JY. The diagnostic values of ultrasound and ultrasound-guided fine needle aspiration in subcentimeter-sized thyroid nodules. Ann Surg Oncol 2012;19:52-9.

17. Kaliszewski K, Diakowska D, Wojtczak B, Migon J, Kasprzyk A, Rudnicki J. The occurrence of and predictive factors for multifocality and bilaterality in patients with papillary thyroid microcarcinoma. Medicine (Baltimore) 2019;98:e15609.

18. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, 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.

19. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Trans Med Imaging 2016;35:1299-312.

20. Ma J, Wu F, Zhu J, Xu D, Kong D. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. Ultrasonics 2017;73:221-30.

21. Ko SY, Lee JH, Yoon JH, Na H, Hong E, Han K, et al. Deep convolutional

neural network for the diagnosis of thyroid nodules on ultrasound. Head Neck 2019;41:885-91.

22. Wang L, Yang S, Yang S, Zhao C, Tian G, Gao Y, et al. Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network. World J Surg Oncol 2019;17:12.

23. Li X, Zhang S, Zhang Q, Wei X, Pan Y, Zhao J, 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.

24. Park VY, Han K, Seong YK, Park MH, Kim EK, Moon HJ, et al. Diagnosis of Thyroid Nodules: Performance of a Deep Learning Convolutional Neural Network Model vs. Radiologists. Sci Rep 2019;9:17843.

25. Koh J, Lee E, Han K, Kim EK, Son EJ, Sohn YM, et al. Diagnosis of thyroid nodules on ultrasonography by a deep convolutional neural network. Sci Rep 2020;10:15245.

26. Kim EK, Park CS, Chung WY, Oh KK, Kim DI, Lee JT, 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.

27. Kwak JY, Han KH, Yoon JH, Moon HJ, Son EJ, Park SH, et al. Thyroid imaging reporting and data system for US features of nodules: a step in establishing better stratification of cancer risk. Radiology 2011;260:892-9.

28. Yoon JH, Lee HS, Kim EK, Moon HJ, Kwak JY. Malignancy Risk Stratification of Thyroid Nodules: Comparison between the Thyroid Imaging Reporting and Data System and the 2014 American Thyroid Association Management Guidelines. Radiology 2016;278:917-24.

29. 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.

30. An C, Choi GH, Lee HS, Kim MJ. Assessment of preoperative magnetic resonance imaging staging in patients with hepatocellular carcinoma undergoing resection compared with the seventh American Joint Committee on Cancer System. Invest Radiol 2012;47:634-41.

31. Marrero JA, Fontana RJ, Barrat A, Askari F, Conjeevaram HS, Su GL, et al. Prognosis of hepatocellular carcinoma: comparison of 7 staging systems in an American cohort. Hepatology 2005;41:707-16.

32. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control 1974;19:716-23.

33. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation 1989;1:541-51.

34. Buda M, Wildman-Tobriner B, Hoang JK, Thayer D, Tessler FN, Middleton WD, et al. Management of Thyroid Nodules Seen on US Images: Deep Learning May Match Performance of Radiologists. Radiology 2019;292:695-701.

35. Kim HL, Ha EJ, Han M. Real-World Performance of Computer-Aided

Diagnosis System for Thyroid Nodules Using Ultrasonography. Ultrasound Med Biol 2019;45:2672-8.

36. Han M, Ha EJ, Park JH. Computer-Aided Diagnostic System for Thyroid Nodules on Ultrasonography: Diagnostic Performance Based on the Thyroid Imaging Reporting and Data System Classification and Dichotomous Outcomes. American Journal of Neuroradiology 2020; doi:10.3174/ajnr.A6922.

37. Yoon JH, Kwon HJ, Kim EK, Moon HJ, Kwak JY. The follicular variant of papillary thyroid carcinoma: characteristics of preoperative ultrasonography and cytology. Ultrasonography 2016;35:47-54.

38. Jeong EY, Kim HL, Ha EJ, Park SY, Cho YJ, Han M. 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.

ABSTRACT (IN KOREAN)

심층 컨볼루션 신경망을 이용한 초음파 상 갑상선 미세결절의 진단

<지도교수 곽 진 영>

연세대학교 대학원 의학과

천 세 현

심층 컨볼루션 신경망 (convolutional neural network, CNN) 을 이용하여 10 mm 미만의 크기를 가지는 갑상선 결절의 진단 및 분류 능력을 평가하고, 이를 영상의학과 의사의 진단 능력과 비교하고자 하였다.

10 mm 초과의 크기를 가지는 13,560 개의 갑상선 결절의 초음파 영상을 이용하여 CNN 을 훈련하였다. 2016년 3월부터 2018년 2월까지 세브란스병원에서 세포흡입검사 혹은 수술적 절제를 통해 양성, 악성 여부가 확인된 10 mm 미만의 크기를 갖는 갑상선 결절 370개를 대상으로 CNN과 영상의학과 의사의 진단 능력을 평가하고, 곡선하면적 (area under curve, AUC), 민감도, 특이도, 정확도 등을 이용하여 비교하였다. 결절의 크기 5 mm 기준으로 하위 그룹을 정의하고 각 그룹에서의 진단 능력을 분석하였다. 또한 갑상선 결절의 악성 위험도에 따른 분류 능력을 서로 비교하였다.

370개의 결절 중 323개가 악성, 47개가 양성이었다. CNN은 영상의학과 의사와 비교하여 유의하게 높은 AUC 값을 보였고 (0.663 vs. 0.567, P=0.045), 더 뛰어난 분류 능력을 보였다. 또한 5 mm 이하 크기를 가지는 결절에 대한 하위분석에서도 CNN은 영상의학과 의사와 비교하여 높은 AUC와 특이도를 보였다.

CNN 은 10 mm 이상의 갑상선 결절의 초음파 영상으로 훈련하였을 때 10 mm 미만의 작은 갑상선 결절을 진단함에 있어 우수한 진단 및 분류 능력을 보인다.

핵심되는 말 : 인공 지능, 갑상선 결절, 초음파