Diagnostic Accuracy Study

Medicine®

OPEN

Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules

Junho Song, MSc[a], Young Jun Chai, MD, PhD[b,*], Hiroo Masuoka, MD, PhD[c], Sun-Won Park, MD, PhD[d], Su-jin Kim, MD, PhD[e], June Young Choi, MD, PhD[f], Hyoun-Joong Kong, PhD[g], Kyu Eun Lee, MD, PhD[e], Joongseek Lee, MArch[a], Nojun Kwak, PhD[a], Ka Hee Yi, MD, PhD[h], Akira Miyauchi, MD, PhD[c]

Abstract

Fine needle aspiration (FNA) is the procedure of choice for evaluating thyroid nodules. It is indicated for nodules >2 cm, even in cases of very low suspicion of malignancy. FNA has associated risks and expenses. In this study, we developed an image analysis model using a deep learning algorithm and evaluated if the algorithm could predict thyroid nodules with benign FNA results.

Ultrasonographic images of thyroid nodules with cytologic or histologic results were retrospectively collected. For algorithm training, 1358 (670 benign, 688 malignant) thyroid nodule images were input into the Inception-V3 network model. The model was pretrained to classify nodules as benign or malignant using the ImageNet database. The diagnostic performance of the algorithm was tested with the prospectively collected internal (n = 55) and external test sets (n = 100).

For the internal test set, 20 of the 21 FNA malignant nodules were correctly classified as malignant by the algorithm (sensitivity, 95.2%); and of the 22 nodules algorithm classified as benign, 21 were FNA benign (negative predictive value [NPV], 95.5%). For the external test set, 47 of the 50 FNA malignant nodules were correctly classified by the algorithm (sensitivity, 94.0%); and of the 31 nodules the algorithm classified as benign, 28 were FNA benign (NPV, 90.3%).

The sensitivity and NPV of the deep learning algorithm shown in this study are promising. Artificial intelligence may assist clinicians to recognize nodules that are likely to be benign and avoid unnecessary FNA.

Abbreviations: DLA = deep learning algorithm, FNA = fine needle aspiration, US = ultrasonography.

Keywords: artificial intelligence, deep learning, thyroid nodule, ultrasound

1. Introduction

Thyroid nodular disease is very common and its prevalence increases with age.[1] Fine needle aspiration (FNA) is the diagnostic procedure of choice. According to the 2015 American Thyroid Association guidelines, FNA is clinically indicated for nodules >1 cm with suspicious features on ultrasonography (US), and for nodules >2 cm, even with a very low suspicion of malignancy.[2] However, 59% to 85% of nodules subject to FNA are shown to be benign and do not require further management.[3,4] In addition, FNA has significant associated risks and medical expenses.[5] Thus, the ability to predict thyroid nodules which are likely to be benign would be beneficial as it would decrease the rate of unnecessary FNA.

Computerized image analysis based on deep learning algorithms (DLAs) has been broadly and rapidly applied across medical fields.[6–10] The process of computerized image analysis is based on artificial neural networks. Artificial neural networks use a multi-step process to automatically learn features of an image, then extract the features, and classify them using an algorithm. This process does not need manual engineering.

Because the thyroid glands are small and located superficially, obtaining representative US images of thyroid nodules is easy, even for inexperienced clinicians. This makes thyroid US images highly suitable for medical image analysis using DLA. Numerous studies analyzed thyroid US images using traditional machine learning or deep learning and showed promising results.[11–13] However, the validity of such publications is limited because the predictive algorithms were trained and tested using radiologists' interpretation, not FNA cytology or surgical pathology. Radiologists' interpretation is often not congruent with FNA cytology or surgical pathology,[14,15] and has inter- and intraobserver variability.

In this study, we investigated the applicability of DLA for the diagnosis of thyroid nodules using US images. We then evaluated the ability of the algorithm to predict benignity or malignancy. For this end, we developed a DLA using US images labeled with

Editor: Gaurav Malhotra.

This work was supported by a multidisciplinary research grant-in-aid from the Seoul Metropolitan Government Seoul National University (SMG-SNU) Boramae Medical Center (02-2018-3, recipient: YJC).

The authors have no conflicts of interest to disclose.

[a] Graduate School of Convergence Science and Technology, Seoul National University, Suwon, [b] Department of Surgery, Seoul Metropolitan Government Seoul National University Boramae Medical Center, Seoul, Korea, [c] Department of Surgery, Kuma Hospital, Kobe, Japan, [d] Department of Radiology, Seoul National University College of Medicine, Seoul Metropolitan Government Seoul National University Boramae Medical Center, Seoul, [e] Department of Surgery, Seoul National University Hospital and College of Medicine, Seoul, [f] Department of Surgery, Seoul National University Bundang Hospital, Seongnam-si, Gyeonggi-do, [g] Department of Biomedical Engineering, Chungnam National University Hospital, Chungnam National University College of Medicine, Daejeon, [h] Department of Internal Medicine, Seoul Metropolitan Government Seoul National University Boramae Medical Center, Seoul, Korea.

[*] Correspondence: Young Jun Chai, Department of Surgery, Seoul Metropolitan Government Seoul National University Boramae Medical Center, 20 Boramaep-ro 5-gil, Dongjak-gu, Seoul 07061, Korea (e-mail: kevinjoon@naver.com).

Medicine (2019) 98:15(e15133)

Received: 26 September 2018 / Received in final form: 18 January 2019 / Accepted: 12 March 2019

http://dx.doi.org/10.1097/MD.0000000000015133

the results of cytologic or histologic results, and evaluated the performance of the algorithm using internal and external test sets.

2. Patients and methods

2.1. Patients

The institutional review board at SMG-SNU Boramae Medical Center, Seoul, Korea, approved this study (L-2017-426). Images of thyroid nodules with FNA cytologic or histologic results obtained from January 2013 to May 2017 were collected consecutively and reviewed. Nodules were selected for FNA according to the Korean Society of Thyroid Radiology (KSThR) FNA guideline recommendations at the time the procedure was performed.[16,17]

Of these cases, between 2013 and 2015, FNA was performed for all suspicious malignant nodules regardless of size, as well as for nodules that were >1 cm even if they seemed benign, according to KSThR guidelines. In 2016, the KSThR adopted the Korean Thyroid Imaging Reporting and Data System (K-TRIADS).[18] Therefore, from 2016, K-TIRADS was used to choose nodules for FNA.

Nodules were labeled as "benign" if their FNA cytology was Bethesda Category II or surgical histology was benign. Nodules were labeled as "malignant" if their FNA cytology was Bethesda Category V/VI or surgical histology was papillary thyroid carcinoma. Nondiagnostic or indeterminate nodules (Bethesda Category I, III, IV) were categorized as benign or malignant according to surgical pathology. They were excluded if not surgically proven to be benign or malignant. A total of 1358 US images (670 benign, 688 malignant) were collected and labeled as benign or malignant (see Images, Supplemental Content, which demonstrates all US images for training and test sets; http://links.lww.com/MD/C929)

2.2. Data preprocessing

The US images containing representative features of thyroid nodules were downloaded in DICOM or TIFF format (Fig. 1). Then the nodules on the images were cropped into squares (299 × 299 pixel) by a single clinician (YJC). We did not implement data augmentation techniques such as flip, changing brightness, or changing scale.

2.3. Training and validation for DLA establishment

Using the preprocessed US images, we trained our neural net to classify benign and malignant nodules. To do so, we used a transfer learning method using the Inception-v3 model, which is the most popular image recognition model and has been previously successfully adapted for medical image analysis.[19–21] The Inception-v3 model was pretrained with >1.2 million images labeled with 1000 semantic classes from the ImageNet Large Scale Visual Recognition Challenge repository.[22] Inception-v3 model architecture consists of the following layers which are pretrained, and contain information that can discriminate between images: a stem layer, 3 × Inception-A layers, 5 × Inception-B layers, 2 × Inception-B layer, a pooling layer, a dropout layer, a fully connected layer, and a softmax layer. For this study, we trained the fully connected layer with 1358 US images to create a new fully connected layer.

We used a "bottleneck layer," with an extremely small number of units (compared with the adjacent layers). A small number of units can aggregate the propagated information and extract fundamental features from the input data.[23] The new fully connected layer was trained with hyperparameters, with a learning rate of 0.01, a batch size of 100, model store frequency of 300, and 7000 training steps. We used validation data, which were 10% of the total training data, in a holdout cross-validation manner. We recorded training accuracy and validation accuracy every 10 training steps for 7000 steps. We identified training step 2100 as the point where the gap between the training accuracy and validation accuracy began to spread. We selected training step 2100 as the final model without overfitting. Benignity or malignancy was presented based on a probability threshold of 0.5.

2.4. US image analysis by radiologists and deep learning algorithm

After the DLA was established, we prospectively collected US images of 1 to 3 cm nodules with FNA cytologic results for an internal test set. The US images were taken using a single system (iU22 system, Philips, Seattle, WA) and reviewed by a single experienced radiologist (SWP). The K-TRIADS was used to evaluate the malignancy risk of each nodule stratified by its US patterns composed of the integrated solidity, echogenicity, and suspicious US features of each nodule.[18] The nodules were categorized as benign (K-TIRADS 2), low suspicion (K-TIRADS 3), intermediate suspicion (K-TIRADS 4), and high suspicion (K-TIRADS 5).

To solve the overfitting problem, we received US images of 1 to 3 cm thyroid nodules from Kuma Hospital, Kobe, Japan, for use as an external test set. Cytologic results were not revealed until the results of the DLA were sent to Kuma Hospital. The US

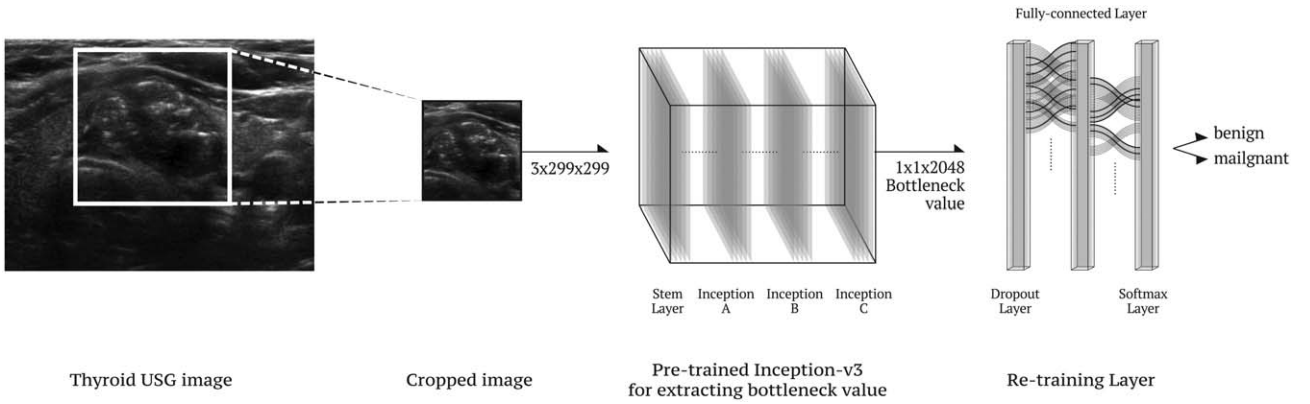

Figure 1. Image analysis process using pre-trained neural network.

Table 1

Internal test set reviewed by radiologist.

K-TIRADS classification by radiologist (n)	FNA cytologic diagnosis (n)
K-TIRADS 2, benign (1)	Benign (34)
K-TIRADS 3, low suspicion (22)	
K-TIRADS 4, intermediate suspicion (8)	
K-TIRADS 5, high suspicion (3)	
K-TIRADS 5, high suspicion (21)	Malignant (21)

FNA = fine needle aspiration, K-TIRADS = the Korean Thyroid Imaging Reporting and Data System.

images were taken with a single system (TUS-A500, Toshiba Medical System, Tokyo, Japan) and reviewed by a single experienced clinician. The nodules were evaluated according to Kuma US classification: the nodules were categorized as benign (class 1–2.5), follicular neoplasm (class 3), and suspected of thyroid carcinoma (class 3.5–5).[24]

To demonstrate the performance of the DLA by proportion of malignancy, the images of the malignant nodules in the external test set were randomly sampled to account for 10%, 20%, 30%, and 40% using random sampling code (https://docs.scipy.org/doc/numpy-1.13.0/reference/generated/numpy.random.randint.html).

3. Results

3.1. Internal test set

The internal test set comprised 55 US nodule images (34 FNA benign, 21 FNA malignant) from SNU-SMG Boramae Medical Center (Table 1). Of the 34 FNA benign nodules, 23 nodules were benign or low suspicion (K-TIRADS 2 or 3) and 11 nodules were intermediate suspicion (K-TIRADS 4 or 5). Three nodules were high suspicion (K-TIRADS 5). All of the 21 FNA malignant nodules were high suspicion (K-TIRADS 5).

Table 2 demonstrates the diagnostic performance of the DLA. Of the 21 FNA malignant nodules, 20 were classified as malignant by the algorithm (sensitivity, 95.2%). Of the 34 FNA benign nodules, 21 were predicted as benign by the algorithm (specificity, 61.8%). Of the 22 nodules that the algorithm classified as benign, 21 were FNA benign (negative predictive value [NPV], 95.5%). Of the 33 nodules algorithm classified as malignant, 20 were FNA malignant (positive predictive value, 60.6%). Figure 2 shows the image of the FNA malignant nodule that was incorrectly classified as benign by the DLA.

Table 2

Diagnostic performance of deep learning algorithm in the internal test set.

	FNA benign (n)	FNA malignant (n)
Benign by algorithm	21	1
Malignant by algorithm	13	20

FNA = fine needle aspiration.

3.2. External test set

The external test set comprised 100 nodules images (50 FNA benign, 50 FNA malignant) from Kuma Hospital (Table 3). All of the FNA benign nodules were benign under Kuma US classification. Of the FNA malignant nodules, 14 were follicular neoplasm and 36 were suspected thyroid carcinoma under Kuma US classification.

Table 4 demonstrates the diagnostic performance of the DLA. Of the 50 FNA malignant nodules, 47 were classified as malignant by the algorithm (sensitivity, 94.0%). Of the 50 FNA benign nodules, 28 were predicted as benign by the algorithm (specificity, 56.0%). Of the 31 nodules that the algorithm classified as benign, 28 were FNA benign (NPV, 90.3%). Of the 69 nodules algorithm classified as malignant, 47 were FNA malignant (positive predictive value, 68.1%). The images of the 3 FNA malignant nodules that were incorrectly classified as benign by the DLA are shown in Figure 3.

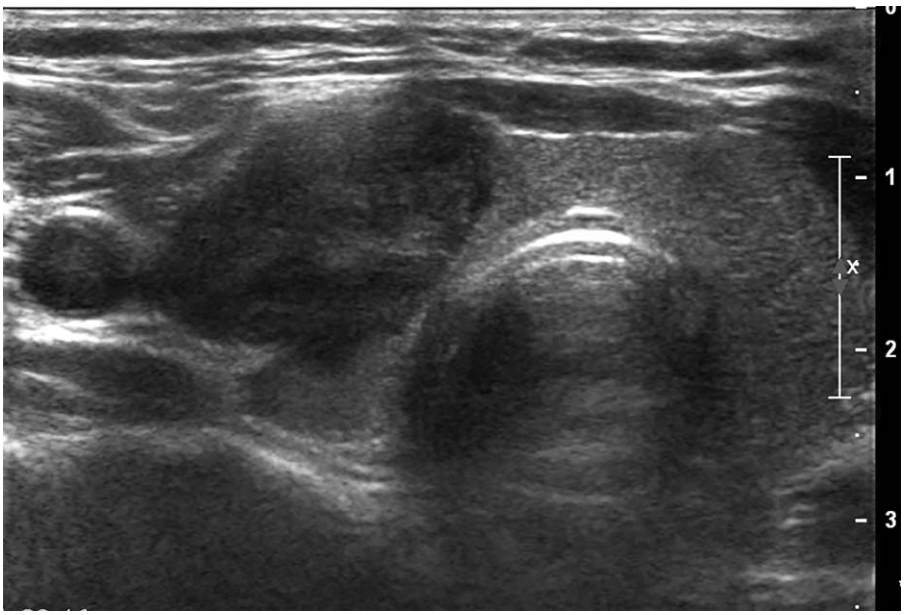

Figure 2. Image of the malignant nodule that was incorrectly classified as benign by the deep learning algorithm in the internal test set.

Table 3

External test set reviewed by radiologist.

Kuma classification (n)	FNA cytologic diagnosis (n)
Benign (50)	Benign (50)
Follicular neoplasm (14)	Malignant (50)
Suspected of thyroid carcinoma (36)	

FNA = fine needle aspiration.

Table 4

Diagnostic performance of deep learning algorithm in the external test set.

	FNA benign (n)	FNA malignant (n)
Benign by algorithm	28	3
Malignant by algorithm	22	47

FNA = fine needle aspiration.

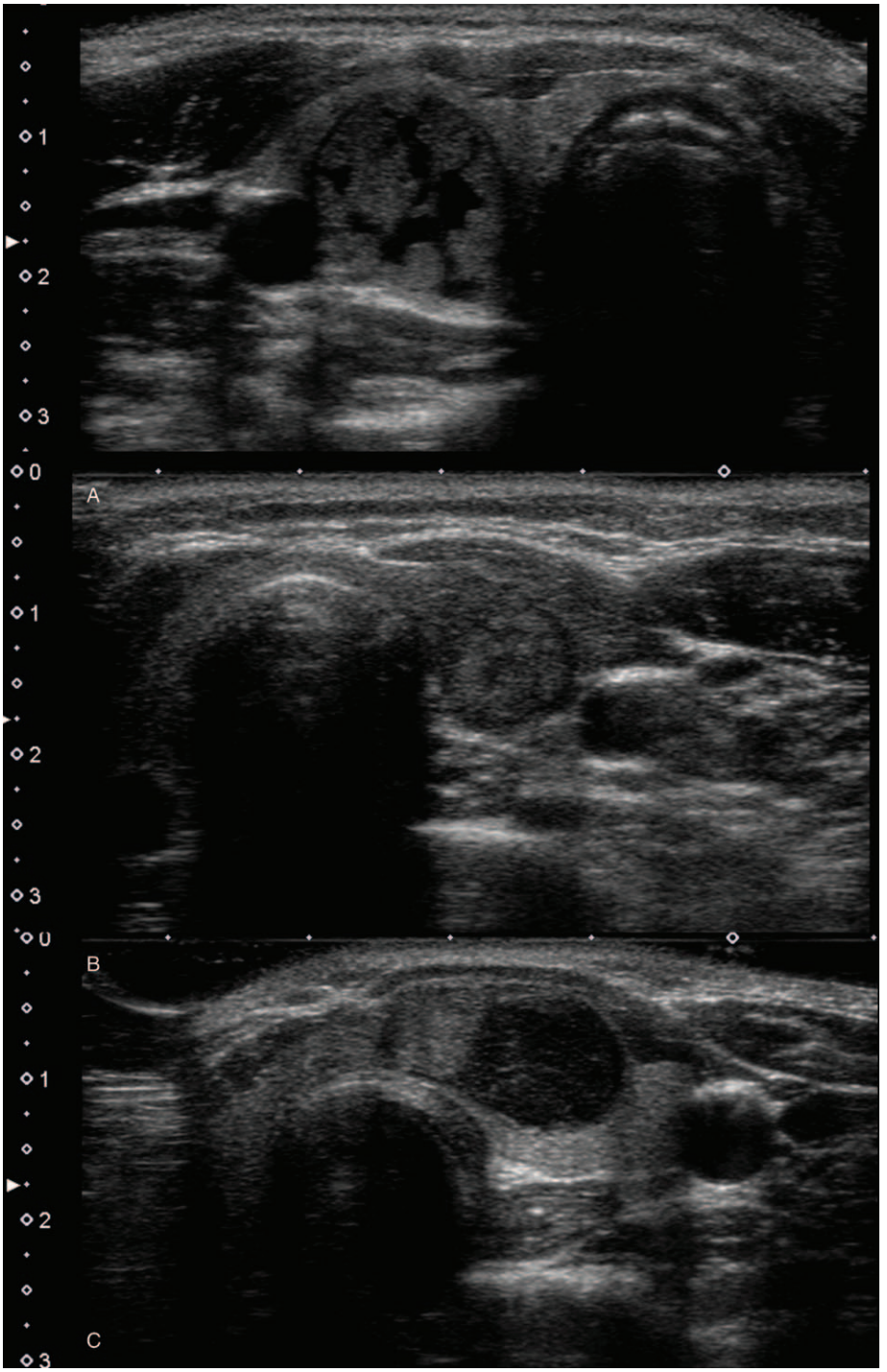

Figure 3. Images of the malignant nodules that were incorrectly classified as benign by the deep learning algorithm in the external test set.

Table 5

Diagnostic performance of deep learning algorithm in the external test set according to the proportion of malignancy.

Proportion of malignant nodules	Sensitivity (%)	Negative predictive value (%)
50% (50 benign, 50 malignant)	94.0	90.3
40% (50 benign, 34 malignant)	91.2	90.3
30% (50 benign, 22 malignant)	90.9	93.3
20% (50 benign, 13 malignant)	92.3	96.6
10% (50 benign, 6 malignant)	100.0	100.0

Table 5 shows the diagnostic performance of the DLA in the external test set by proportion of malignancy. The sensitivity ranged from 94.0% to 100%, and NPV from 90.3% to 100%.

4. Discussion

In this study, the thyroid nodules classified as benign by the DLA were highly likely to be FNA benign. FNA is the most reliable and cost-effective diagnostic test for thyroid nodules. However, complications that can accompany FNA cannot be ignored. Common complications include pain (88%–92% of patients)[25,26] and blood extravasation (1.9%–6.4% of patients).[27] Rare, but more serious complications include recurrent laryngeal nerve palsy,[28] vasovagal reaction,[29] and potentially life-threatening airway obstruction.[27] Therefore, the ability to screen nodules that may not require FNA carries significant clinical implications.

Misdiagnosis of true malignancy as benign (false-negative) may cause delay in surgical intervention.[30] Misdiagnosis of true benign as malignant (false-positive) would merely lead to FNA, which is already indicated in cases in this study. Thus, sensitivity and NPV are the 2 most meaningful parameters to evaluate the diagnostic performance of the DLA. Image quality can be affected by race, underlying thyroiditis, or equipment. The sensitivity and NPV of a DLA can decrease if the quality of the images differs from that of the training set. Moreover, the sensitivity and NPV can decrease with increasing disease prevalence.[31] To evaluate the influence of the 2 possible variables (image quality and increasing prevalence), we assessed the performance of the algorithm at various proportions of malignancy and used an external validation test set from an institute in a different country.

For the external test set, the sensitivity of the algorithm ranged from 91.2% to 100%, and NPV ranged from 90.3% to 100%, according to the proportion of malignancy. In the present study, the sensitivity and NPV of the DLA were 100%, based on the assumption that the prevalence of malignancy (Bethesda Category V/VI) in the external test set was similar to the 9.3% reported in literature.[3]

The diagnostic accuracy was comparable between image analysis technology and radiologists in a previous study.[11] The study trained a program using radiologists' diagnoses, and can only predict US category, not cytologic or histologic results. The clinical significance of such models may be considered limited because FNA of the nodules would still be necessary if indicated. In contrast, we developed a DLA that provides clinical significance in terms of FNA decision-making because it was trained on cytologic or histologic test results. The high NPV of the DLA in the present study suggests that this technology may facilitate clinical decisions and avoid unnecessary FNA. The clinical significance of the DLA of this study is higher when we consider that the nodules in the test sets had all undergone FNA. Although the DLA would not replace the clinician, it could certainly serve as a clinical decision support tool, especially if an experienced clinician was unavailable.

Image databases have been established to support the growing need for big data. Specific groups of images, such as thyroid US, are being accumulated for general use.[32] In addition, online deep learning sources for developing DLAs are now publicly available. Such deep learning libraries and image databases enable researchers to develop DLAs for image analysis, provided that they have sufficient number of images to train their algorithm and are able to perform hyperparameter optimization experiments. Assuming that each of the US images is prepared (downloaded in the appropriate format and cropped to size), training a new DLN using this method takes around 30 hours. Generally, to find the most appropriate DLN model, the parameters are tested and adjusted repeatedly according to the researcher's needs. In the present study, we repeated DLN training 3 times, which took 10 days in total. Once the DLA was established, the benignity or malignancy of a given nodule was able to be predicted in almost real time (0.07 seconds).

The DLA developed in this study was pretrained with the ImageNet database, and retrained with 1138 US images from the researchers' institute. The ImageNet database contains over 1.2 million images of things commonly seen in daily life (not medical or US images). The images are labeled using 1000 semantic classes. Models can be trained using huge quantities of image data and that knowledge can be transferred to the medical image domain. We found it more effective to pretrain our DLA using the massive ImageNet database than to use a limited number of medical images.[33] Clearly, future DLAs may be better developed using a large database of US images for training.

For image preparation, we simply cropped the nodule part of the images into a square. This process did not require any specialized skills such as segmentation or demarcation. Data augmentation is a technique used to increase the number of training data artificially by changing the ratio of width to height, changing colors, or using horizontal flip. It is reported to be an essential technique required by DLAs to achieve good performance.[34] However, we did not apply data augmentation because it has a high potential to distort shape, margin, echogenicity, and calcification, which are essential for differentiating benignity and malignancy.

Despite a small training set and the simplicity of image preparation, the DLA in this study displayed excellent performance in selecting benign nodules. In general, the performance or accuracy of a DLA increases logarithmically based on the quantity of data used for training,[35] and thus the number of training images is more important than their quality.[36] In this sense, the DLA for the thyroid US diagnosis is expandable because the algorithm can be easily shared, and it will become more accurate as it is trained further with more images. Moreover, the sensitivity and NPV of the external test set were comparable to those of the internal test set, despite the fact that the study population, image quality, and the equipment used were different between institutes. This may be because thyroid nodules show relatively consistent and typical characteristics in shape, margin, echogenicity, and calcification. This suggests that further advancements in the accuracy of the DLA are possible by expanding the US image dataset for training.

This study has limitations. First, we considered Bethesda Category V and VI to be malignant. However, in some cases, category V and even category VI may be "true benign" based on

surgical histology. Although none of the category V or VI nodules in this study were shown to be benign on surgery (data not shown), it should be taken into account that the malignancy proportion of Bethesda category V and VI varies between institutes. Second, some FNA benign nodules were classified by the DLA as malignant (low specificity). Lower specificity in image analysis by DLA is often observed.[37] Inferred from the results, the DLA in this study detected malignant features of nodules sensitively and classified them as malignant if any of those features were suspicious. Considering that the role of DLA still has limited application in assisting clinicians to select benign nodules without overlooking malignant nodules, high sensitivity and NPV are more important than specificity. Better DLAs are continuously being developed because DLA technology is rapidly evolving. In the near future, DLAs are poised to become more accurate and more helpful for clinicians as a larger volume of images is accumulated.

5. Conclusions

We developed a DLA for the analysis of thyroid US images. Although highly experienced clinicians outperform DLA, the sensitivity and NPV of the DLA in this study are promising. Applying artificial intelligence for the evaluation of thyroid nodules may help clinicians to reduce the number of unnecessary FNAs, in the near future. Such image analysis models are likely to be broadly adopted after their applicability has been demonstrated in larger series.

Author contributions

Conceptualization: Young Jun Chai.
Data curation: Junho Song, Hiroo Masuoka, Sun-Won Park.
Funding acquisition: Young Jun Chai.
Methodology: Hiroo Masuoka, Akira Miyauchi.
Resources: Hyoun-Joong Kong.
Software: Hyoun-Joong Kong.
Supervision: Su-jin Kim, June Young Choi, Kyu Eun Lee, Joongseek Lee, Nojun Kwak, Ka Hee Yi.
Validation: Sun-Won Park, Akira Miyauchi.
Writing – original draft: Young Jun Chai.
Young Jun Chai orcid: 0000-0001-8830-3433.

References

[1] Mazzaferri EL. Management of a solitary thyroid nodule. N Engl J Med 1993;328:553–9.
[2] 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;26:1–33.
[3] Jo VY, Stelow EB, Dustin SM, et al. Malignancy risk for fine-needle aspiration of thyroid lesions according to the Bethesda System for Reporting Thyroid Cytopathology. Am J Clin Pathol 2010;134:450–6.
[4] Harvey AM, Mody DR, Amrikachi M. Thyroid fine-needle aspiration reporting rates and outcomes before and after Bethesda implementation within a combined academic and community hospital system. Arch Pathol Lab Med 2013;137:1664–8.
[5] Cesareo R, Naciu A, Barberi A, et al. A rare and severe complication following thyroid fine needle aspiration: retropharyngeal cellulitis. Int J Endocrinol Metab 2016;14:e39174.
[6] Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017;318:2211–23.
[7] Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318:2199–210.
[8] Becker AS, Mueller M, Stoffel E, et al. Classification of breast cancer from ultrasound imaging using a generic deep learning analysis software: a pilot study. Br J Radiol 2017;20170576.
[9] Zreik M, Lessmann N, van Hamersvelt RW, et al. Deep learning analysis of the myocardium in coronary CT angiography for identification of patients with functionally significant coronary artery stenosis. Med Image Anal 2017;44:72–85.
[10] Choi JW, Ku Y, Yoo BW, et al. White blood cell differential count of maturation stages in bone marrow smear using dual-stage convolutional neural networks. PLoS One 2017;12:e0189259.
[11] 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.
[12] Chi J, Walia E, Babyn P, et al. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. J Digit Imaging 2017;30:477–86.
[13] 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.
[14] Brito JP, Gionfriddo MR, Al Nofal A, et al. The accuracy of thyroid nodule ultrasound to predict thyroid cancer: systematic review and meta-analysis. J Clin Endocrinol Metab 2014;99:1253–63.
[15] Wei X, Li Y, Zhang S, et al. Meta-analysis of thyroid imaging reporting and data system in the ultrasonographic diagnosis of 10,437 thyroid nodules. Head Neck 2016;38:309–15.
[16] Moon WJ, Baek JH, Jung SL, et al. Ultrasonography and the ultrasound-based management of thyroid nodules: consensus statement and recommendations. Korean J Radiol 2011;12:1–4.
[17] Shin JH, Baek JH, Chung J, 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.
[18] Na DG, Baek JH, Sung JY, et al. Thyroid imaging reporting and data system risk stratification of thyroid nodules: categorization based on solidity and echogenicity. Thyroid 2016;26:562–72.
[19] Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559–67.
[20] Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8.
[21] Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402–10.
[22] Szegedy C, Vanhoucke V, loffe S, Shlens J, Z W. Rethinking the Inception Architecture for Computer Vision. arXiv preprint arXiv:1512.00567v3. 2015.
[23] Nakashika T, Yoshioka T, Takiguchi T, et al. Convolutive bottleneck network with dropout for dysarthric speech recognition. Trans Mach Learn Artif Intell 2014;2:1–5.
[24] Ito Y, Amino N, Miyauchi A. Thyroid ultrasonography. World J Surg 2010;34:1171–80.
[25] Gursoy A, Ertugrul DT, Sahin M, et al. Needle-free delivery of lidocaine for reducing the pain associated with the fine-needle aspiration biopsy of thyroid nodules: time-saving and efficacious procedure. Thyroid 2007;17:317–21.
[26] Gursoy A, Ertugrul DT, Sahin M, et al. The analgesic efficacy of lidocaine/prilocaine (EMLA) cream during fine-needle aspiration biopsy of thyroid nodules. Clin Endocrinol (Oxf) 2007;66:691–4.
[27] Polyzos SA, Anastasilakis AD. Systematic review of cases reporting blood extravasation-related complications after thyroid fine-needle biopsy. J Otolaryngol Head Neck Surg 2010;39:532–41.
[28] Tomoda C, Takamura Y, Ito Y, et al. Transient vocal cord paralysis after fine-needle aspiration biopsy of thyroid tumor. Thyroid 2006;16:697–9.
[29] Khoo TK, Baker CH, Hallanger-Johnson J, et al. Comparison of ultrasound-guided fine-needle aspiration biopsy with core-needle biopsy in the evaluation of thyroid nodules. Endocr Pract 2008;14:426–31.

[30] Chai YJ, Suh H, Yi JW, et al. Factors associated with the sensitivity of fine-needle aspiration cytology for the diagnosis of follicular variant papillary thyroid carcinoma. Head Neck 2016;38(Suppl. 1):E1467–1471.
[31] Leeflang MM, Rutjes AW, Reitsma JB, et al. Variation of a test's sensitivity and specificity with disease prevalence. CMAJ 2013;185: E537–44.
[32] Pedraza L, Vargas C, Narváez F, Durán O, Muñoz E, E R. An open access thyroid ultrasound image database. Proc. SPIE 9287, 10th International Symposium on Medical Information Processing and Analysis, 92870W. 2015.
[33] Tajbakhsh N, Shin JY, Gurudu SR, et al. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? arXiv preprint arXiv:1706.00712v1. 2017.
[34] Sajjadi M, Javanmardi M, T T. Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning. arXiv preprint arXiv:1606.04586v1. 2016.
[35] Sun C, Shrivastava A, Singh S, A G. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. arXiv preprint arXiv: 1707.02968. 2017.
[36] Rolnick D, Veit A, Belongie S, N S. Deep Learning is Robust to Massive Label Noise. arXiv preprint arXiv:1705.10694v2. 2017.
[37] Han SS, Park GH, Lim W, et al. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network. PLoS One 2018;13:e0191493.