Artificial intelligence for thyroid nodule ultrasound image analysis

Young Jun Chai[1#], Junho Song[2#], Mohammad Shaear[3], Ka Hee Yi[4]

[1]Department of Surgery, Seoul Metropolitan Government Seoul National University Boramae Medical Center, Seoul, Korea; [2]Graduate School, Convergence Science and Technology Seoul National University, Suwon, Korea; [3]Head and Neck Endocrine Surgery Division, Department of Otolaryngology-Head and Neck Surgery, The Johns Hopkins Hospital, Baltimore, MD, USA; [4]Department of Internal Medicine, Seoul Metropolitan Government Seoul National University Boramae Medical Center, Seoul, Korea

Contributions: (I) Conception and design: YJ Chai; (II) Administrative support: KH Yi; (III) Provision of study materials or patients: J Song; (IV) Collection and assembly of data: YJ Chai; (V) Data analysis and interpretation: YJ Chai; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Young Jun Chai, MD, PhD. Department of Surgery, Seoul Metropolitan Government - Seoul National University Boramae Medical Center, 20 Boramae-ro 5-gil, Dongjak-gu, Seoul, 156-70, Korea. Email: kevinjoon@naver.com.

Abstract: Deep learning (DL) as part of artificial intelligence (AI) is based on artificial neural networks, which use a multi-step process to automatically analyze features of an image, then classify them. Medical image analysis using DL has been rapidly adopted across multiple medical fields. Neck ultrasound (US) is a gold standard diagnostic modality in thyroid nodules, and suitable for DL diagnosis because the characteristics of the thyroid nodules can be captured in one representative image. Therefore, a number of studies applied DL in analyzing neck US and tried to predict malignancy risk of the thyroid nodules, and showed relatively high diagnostic performances. For successful DL analysis for the thyroid US images, several issues should be solved which includes image selection by experienced clinicians, proper quality of US, enough number of US images to train the DL, accurate labeling, and adequate hyperparameters. However, DL analysis is still in its early stages, and the questions are unanswered yet. We do not know how many images we need to obtain for proper training, and there is no standard of US quality. Moreover, DL analysis may not be able to classify indeterminate nodules into benign or malignant even in the near future, not to mention now. In this review, we will discuss the current status, limitations, and the future directions of DL for thyroid US image analysis from a clinician's point of view.

Keywords: Thyroid gland; ultrasound (US); artificial intelligence (AI); deep learning (DL)

Received: 08 October 2019. Accepted: 30 March 2020; Published: 30 June 2020.
doi: 10.21037/aot.2020.04.01
View this article at: http://dx.doi.org/10.21037/aot.2020.04.01

Why are ultrasound (US) images of thyroid nodules suitable for deep learning (DL) analysis?

DL is a part of artificial intelligence (AI) systems, which is designed to have human's way of thinking. DL has been applied in medical image analysis such as chest radiographs, retinal image, pathology, and US images, and showed comparable diagnostic performance with clinicians (1-4). To allow the DL to analyze images, we should input hundreds or thousands of images labelled with the answers, which is called training. Then the DL algorithm trains itself by extracting specific features from the images, and becomes able to predict the answer when a new test image is given. For instance, if one properly trains a DL algorithm with multiple images of cats and dogs labelled with the answers, the DL extracts the features of the two species and finally becomes to be able to differentiate them.

Neck US is a safe diagnostic imaging modality and the gold standard method for evaluation of thyroid nodules (5). Moreover, recent high-resolution US is easily utilized by clinicians for obtaining and interpreting sonographic neck images including thyroid nodules. The characteristics

Table 1 Ultrasonic Features suggestive of malignant thyroid nodules

Hypoechoic nodule
Taller than wide shape of the nodule
Irregular margin of the nodules
Presence of microcalcifications

of these nodules can be captured in one representative image making thyroid US suitable for DL analysis using convolutional neural network. This can be established by adding the features of thyroid nodules obtained by US to a DL algorithm in the form of a whole captured image or alternatively by cropping the image into multiple squares. The latter is generally preferred to avoid the influence of neighboring structures such as trachea, carotid artery, or muscles on the efficacy and accuracy of a DL algorithm.

Previous studies using Computer-aided diagnosis (CAD) for thyroid nodule US images

CAD system is based on classical machine learning. Unlike DL which itself determines and extracts key features from the images, CAD system requires that human should define the key features of a subject on which the prediction is based. In the CAD system for thyroid US, human should define the malignant features of the thyroid nodules such as irregular margin, taller-than-wide shape, markedly hypoechoic echogenicity, and presence of calcification to allow the CAD to predict malignant nodules. Then, the CAD quantifies each predetermined feature and eventually predicts if the given nodule is benign or malignant.

Choi *et al*. (6) used commercialized CAD system for the diagnosis of US images of thyroid nodules. They reported that CAD system showed a similar sensitivity as the experienced radiologist (90.7% *vs.* 88.4%) and lower specificity (74.6% *vs.* 94.9%). Similarly, Jeong *et al*. (7) used the same commercialized CAD system and reported that the CAD system had comparable sensitivity and lower specificity than experienced radiologist. The reason for the relatively high sensitivity and low specificity could be because the CAD system is more sensitive and consistent than the human in identifying the malignant features of thyroid nodules (*Table 1*).

Previous studies using DL for thyroid nodule US images

Unlike CAD, DL does not require predetermination of malignant features but instead the final result of the thyroid biopsy or the specimen. After introducing these results in a DL algorithm, DL works to independently determine the different radiographic features (unbeknownst to the clinician) that are used to interpret future US images. Interestingly, these features may not include the standard factures we use for US diagnosis (e.g., size, shape, etc.). Furthermore, we cannot know the features that DL uses for training or prediction, and thus the DL algorithm is referred to as a black box. DL algorithms generally outperform CAD systems, and the majority of recent studies use DL.

Ko *et al*. (8) used 439 US images for the training set and tested 150 images, and reported that DL algorithm showed comparable performance with radiologists (AUC 0.834 to 0.850 for DL and 0.805 to 0.860 for radiologist). Moreover, Song *et al*. (9) used 1,358 images from a training dataset and tested an algorithm with an internal (n=55) and external (n=100) test set. The sensitivity for the internal and external test set was 95.2% and 94.0% respectively. Buda *et al*. (10) used images from 1,278 and 99 nodules as a training and test set, respectively. They reported 87% sensitivity and 52% specificity in recommending further intervention, which was comparable with that recommended by radiologists.

How do we obtain enough images for successful DL algorithm training?

To train the DL algorithm using US images only, which is called training from *scratch*, a large amount of labeled US images are necessary because the diagnostic performance of a DL algorithm improves according to the size of the training dataset (11). However, the amount of collectable data is limited due to manpower and costs restrictions. In addition, it is unknown how many images are required for successful training. There are several methods however to address this limitation. One of the popular methods is *transfer learning* which saves time as it uses a pre-trained model. A pre-trained model is trained on a large benchmark dataset to solve a problem similar to the one that we want to solve. For instance, *Inception* is one of the most popular models, and pre-trained with the ImageNet database, which contains over 1.2 million images of commonly seen items in daily life. Using a pre-trained model is more efficient than training the whole layer of the DL algorithm despite the dataset not including medical or includes US images (12).

Another method is data augmentation. Data augmentation generates more images artificially by changing the ratio of width to height, adding noise, changing colors, or using horizontal flip. It is reported to be helpful to achieve better DL performance (13). Although data augmentation is useful to increase the size of a training set and has been used in thyroid nodule image analysis (14), caution should be taken because it has a high potential to distort shape, margin, echogenicity, and calcification, which are essential elements for sonographic diagnosis of thyroid nodules (15).

Limitations of DL analysis for thyroid nodule US images

Limitations in US image collection

US is a convenient and reliable diagnostic tool when evaluating thyroid nodules. In general, one representative image contains enough information to delineate the nature of the nodule. Therefore, US fits well with the concept of DL. Regardless, there are high intra- and inter-reader variability in US image acquisition, and there is still a chance that the captured image of a thyroid nodule may not completely represent the lesion. For example, the features suggesting malignancy such as micro-calcification or irregular margin may not be adequately captured, or some features may look differently between axial and transverse images. This limitation might decrease the accuracy of US and subsequently the performance of DL.

Indeterminate category

American College of Radiology Thyroid Imaging Reporting And Data System (ACS TIRADS) is a risk stratification system evaluating thyroid nodules (16). The risk of malignancy is determined by five categories including composition, echogenicity, shape, margin, extra-thyroidal extension, and echogenic foci on US. Yet, ACS TIRADS is not confirmative, and FNAC is recommended for further evaluation of radiologically suspicious nodules. However, some FNAC results, which are reported as Bethesda categories, are still not confirmative (17). Category III/IV/V nodules can have diverse results such as benign, follicular thyroid carcinoma (FTC), variant type papillary thyroid carcinoma (PTC), or PTC on surgical pathology.

DL analysis of thyroid nodules are solely based on US findings. The accuracy of the DL is greatly influenced by the proportion of the nodules with indeterminate categories, which consists of 16% to 38% of the FNAC-tested nodules (18). FTC is usually diagnosed as indeterminate or benign category on FNAC, and has more benign feature on US than PTC. The more FTC is included in the dataset, the worse the diagnostic performance should be. Nonetheless, the researchers can increase the diagnostic performance of DL algorithm by excluding the indeterminate category as well as FTCs in the training or test set. However, the results cannot be applicable to real practice. This is why the number of indeterminate nodules on FNAC such as FTC used for training and testing must be mentioned in studies.

To illustrate this, a recently published paper reported that DL was trained with more than 300,000 thyroid US images, and showed relatively high specificity and sensitivity (14). However, of the 17,627 malignant nodules they used for training set, only 74 (0.4%) were FTCs, and 17,440 (98.9%) were PTC. Moreover, the number of FTC used for the test set was only 4 out of 1,194 nodules. Considering the incidence of FTC is one seventh of that of PTC (19), the results in this study should be interpreted carefully.

Likewise, variant types of PTC such as follicular variant PTC, which accounts for 12% to 30% of all PTC, also have less malignant US features compared to classical PTC (20,21). The more variant types of PTC are included in the training or test set, the wores the diagnostic performance of DL becomes. Therefore, care must be taken when interpreting study results in which most of the included malignant nodules are classic PTC (8).

Additionally, the gold standard in diagnosing a thyroid nodule is based on surgical pathology. However, some nodules are not removed by surgery, and the ultimate diagnosis is based on FNAC, with its inherent limitations mentioned above (9,22).

Nodules with indeterminate category on FNAC should undergo further evaluation with molecular testing or be removed surgically. Therefore, it could be more practical to train DL to discriminate thyroid nodules into groups that require surgery versus those that do not rather than into benign or malignant. Further research however is needed to address these concerns.

Further practical considerations

The results of DL analysis are presented as probability of benignity or malignancy ranging 0 to 1, not simply as benign or malignant. In general, benign or malignant results are presented based on a probability threshold of 0.5. However, the threshold can be arbitrarily adjusted as needed, and the sensitivity and the specificity to predict malignancy change depending on the threshold. If the threshold is set to increase the specificity, the chance for misinterpreting true malignant nodules as benign, resulting in undertreatment of malignant nodules may increase.

Therefore, the threshold should be set to have high sensitivity in order that the false negative rate (predicting malignant as benign) can be minimized even if the false positive rate (predicting benign as malignant) is somewhat high. Considering more than 90% of the thyroid nodules which undergo FNAC turn out to be benign (23), and misdiagnosis of true benign as malignant would merely lead to FNAC, this scenario is desirable because DL can help patients to avoid unnecessary FNAC. Using DL as a decision support tool can be considered if an experienced clinician is not available.

Lastly, DL algorithms have been trained and tested using one 2D US image from each nodule. However, thyroid nodules are three-dimensional, and one representative US image may not completely reflect all pertinent features. DL analysis using multiple US images should be tried in the future to increase diagnostic capability.

Conclusions

Applying DL in the diagnosis of thyroid nodules is still in the developmental stage. There are several limitations related to efficient collection of US images, setting a proper threshold for predicting malignancy, and proper inclusion of indeterminate nodules into the dataset. Although currently developed DL cannot replace standard practice in the diagnosis of thyroid nodules, it might serve as an adjunctive tool to support the decision-making process for biopsy and surgery in the future.

Acknowledgments

Funding: 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).

Footnote

Provenance and Peer Review: This article was commissioned by the Guest Editors (Jonathon Russell and Jeremy Richmon) for the series “The Management of Thyroid Tumors in 2021 and Beyond” published in *Annals of Thyroid*. The article has undergone external peer review.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/aot.2020.04.01). The series “The Management of Thyroid Tumors in 2021 and Beyond” was commissioned by the editorial office without any funding or sponsorship. The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Heo SJ, Kim Y, Yun S, et al. Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in Chest Radiographs in Annual Workers' Health Examination Data. International journal of environmental research and public health 2019;16:250.
2. 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.
3. Becker AS, Mueller M, Stoffel E, et al. Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. Br J Radiol 2018;91:20170576.
4. 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.
5. Haugen BR, Alexander EK, Bible KC, et al. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid

Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid 2016;26:1-133.

6. Choi YJ, Baek JH, Park HS, et al. A Computer-Aided Diagnosis System Using Artificial Intelligence for the Diagnosis and Characterization of Thyroid Nodules on Ultrasound: Initial Clinical Assessment. Thyroid 2017;27:546-52.
7. Jeong EY, Kim HL, Ha EJ, et al. Computer-aided diagnosis system for thyroid nodules on ultrasonography: diagnostic performance and reproducibility based on the experience level of operators. Eur Radiol 2019;29:1978-85.
8. Ko SY, Lee JH, Yoon JH, et al. Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound. Head Neck 2019;41:885-91.
9. Song J, Chai YJ, Masuoka H, et al. Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules. Medicine (Baltimore) 2019;98:e15133.
10. Buda M, Wildman-Tobriner B, Hoang JK, et al. Management of Thyroid Nodules Seen on US Images: Deep Learning May Match Performance of Radiologists. Radiology 2019;292:695-701.
11. Sun C, Shrivastava A, Singh S, et al. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. 2017. arXiv:170702968.
12. Tajbakhsh N, Shin JY, Gurudu SR, et al. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Trans Med Imaging 2016;35:1299-312.
13. Sajjadi M, Javanmardi M, Tasdizen T. Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning. 2016. arXiv:1606.04586.
14. Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol 2019;20:193-201.
15. Akkus Z, Cai J, Boonrod A, et al. A Survey of Deep-Learning Applications in Ultrasound: Artificial Intelligence-Powered Ultrasound for Improving Clinical Workflow. J Am Coll Radiol 2019;16:1318-28.
16. Tessler FN, Middleton WD, Grant EG, et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. J Am Coll Radiol 2017;14:587-95.
17. Cibas ES, Ali SZ. The Bethesda System for Reporting Thyroid Cytopathology. Thyroid 2009;19:1159-65.
18. Bongiovanni M, Spitale A, Faquin WC, et al. The Bethesda System for Reporting Thyroid Cytopathology: a meta-analysis. Acta Cytol 2012;56:333-9.
19. Aschebrook-Kilfoy B, Grogan RH, Ward MH, et al. Follicular thyroid cancer incidence patterns in the United States, 1980-2009. Thyroid 2013;23:1015-21.
20. Kim DS, Kim JH, Na DG, et al. Sonographic features of follicular variant papillary thyroid carcinomas in comparison with conventional papillary thyroid carcinomas. J Ultrasound Med 2009;28:1685-92.
21. Chai YJ, Kim SJ, Kim SC, et al. BRAF mutation in follicular variant of papillary thyroid carcinoma is associated with unfavourable clinicopathological characteristics and malignant features on ultrasonography. Clin Endocrinol (Oxf) 2014;81:432-9.
22. Ma J, Wu F, Zhu J, et al. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. Ultrasonics 2017;73:221-30.
23. Frates MC, Benson CB, Charboneau JW, et al. Management of thyroid nodules detected at US: Society of Radiologists in Ultrasound consensus conference statement. Ultrasound Q 2006;22:231-8; discussion 239-40.

doi: 10.21037/aot.2020.04.01

Cite this article as: Chai YJ, Song J, Shaear M, Yi KH. Artificial intelligence for thyroid nodule ultrasound image analysis. Ann Thyroid 2020;5:8.