


Management of Thyroid Nodules Seen on US Images: Deep Learning May Match Performance of Radiologists

Mateusz Buda, MSc • Benjamin Wildman-Tobriner, MD • Jenny K. Hoang, MBBS, MHS • David Thayer, PhD, MD • Franklin N. Tessler, MD • William D. Middleton, MD • Maciej A. Mazurowski, PhD

From the Department of Radiology, Duke University School of Medicine, 2424 Erwin Road, Suite 302, Durham, NC 27705 (M.B., B.W.T., J.K.H., M.A.M.); Mallinckrodt Institute of Radiology, Washington University School of Medicine, St Louis, Mo (D.T., W.D.M.); Department of Radiology, University of Alabama at Birmingham, Birmingham, Ala (F.N.T.); and Department of Electrical and Computer Engineering, Duke University, Durham, NC (M.A.M.). Received June 5, 2018; revision requested July 26; revision received April 23, 2019; accepted May 29. **Address correspondence** to M.B. (e-mail: mateusz.buda@duke.edu).

M.B., B.W.T., J.K.H., and M.A.M. supported by the Putman Innovation Award.

Conflicts of interest are listed at the end of this article.

Radiology 2019; 292:695–701 • <https://doi.org/10.1148/radiol.2019181343> • Content codes: 

Background: Management of thyroid nodules may be inconsistent between different observers and time consuming for radiologists. An artificial intelligence system that uses deep learning may improve radiology workflow for management of thyroid nodules.

Purpose: To develop a deep learning algorithm that uses thyroid US images to decide whether a thyroid nodule should undergo a biopsy and to compare the performance of the algorithm with the performance of radiologists who adhere to American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS).

Materials and Methods: In this retrospective analysis, studies in patients referred for US with subsequent fine-needle aspiration or with surgical histologic analysis used as the standard were evaluated. The study period was from August 2006 to May 2010. A multitask deep convolutional neural network was trained to provide biopsy recommendations for thyroid nodules on the basis of two orthogonal US images as the input. In the training phase, the deep learning algorithm was first evaluated by using 10-fold cross-validation. Internal validation was then performed on an independent set of 99 consecutive nodules. The sensitivity and specificity of the algorithm were compared with a consensus of three ACR TI-RADS committee experts and nine other radiologists, all of whom interpreted thyroid US images in clinical practice.

Results: Included were 1377 thyroid nodules in 1230 patients with complete imaging data and conclusive cytologic or histologic diagnoses. For the 99 test nodules, the proposed deep learning algorithm achieved 13 of 15 (87%; 95% confidence interval [CI]: 67%, 100%) sensitivity, the same as expert consensus ($P > .99$) and higher than five of nine radiologists. The specificity of the deep learning algorithm was 44 of 84 (52%; 95% CI: 42%, 62%), which was similar to expert consensus (43 of 84; 51%; 95% CI: 41%, 62%; $P = .91$) and higher than seven of nine other radiologists. The mean sensitivity and specificity for the nine radiologists was 83% (95% CI: 64%, 98%) and 48% (95% CI: 37%, 59%), respectively.

Conclusion: Sensitivity and specificity of a deep learning algorithm for thyroid nodule biopsy recommendations was similar to that of expert radiologists who used American College of Radiology Thyroid Imaging and Reporting Data System guidelines.

© RSNA, 2019

Online supplemental material is available for this article.

Imaging with US remains an accurate method to guide recommendation for management of thyroid nodules (1), although interpretation variability and overdiagnosis represent continual challenges (2,3). To help radiologists improve consistency, several organizations have developed imaging criteria to aid in the selection of nodules recommended for fine-needle aspiration (FNA) biopsy. In 2017, the American College of Radiology (ACR) published its Thyroid Imaging Reporting and Data System (TI-RADS) (4). Similar to its predecessors, ACR TI-RADS is on the basis of US features and maximum nodule size. ACR TI-RADS has been shown to increase accuracy and specificity compared with other systems (5), enhance report quality, and improve recommendations for management (6).

Despite these potential benefits, certain barriers may prevent radiologists from adopting or using ACR TI-RADS. First, a high interobserver variability among radiologists'

interpretations has been shown with the system ($\kappa = 0.51$) (2). Such variability may lead to inconsistent recommendations for nodule management between readers. Second, evaluating multiple nodules (with multiple features per nodule) can be labor intensive and could be more time consuming for some radiologists. Any practice that adds time to an already busy radiology workflow could serve as a disincentive for adopting best practices.

Because of these types of challenges, the medical community has started to use deep learning (7). Deep learning represents an approach to artificial intelligence that has been increasingly applied throughout medicine, with emerging applications in fields such as dermatology (8), ophthalmology (9), and radiology (10,11). Recent deep learning research in radiology has shown algorithm performance comparable to radiologists (12), and as the field continues to grow the variety and number of possible uses for deep

Abbreviations

ACR = American College of Radiology, AUC = area under the receiver operating characteristic curve, CI = confidence interval, FNA = fine-needle aspiration, TI-RADS = Thyroid Imaging Reporting and Data System

Summary

A deep convolutional neural network that uses American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) features for training achieved similar sensitivity and specificity for recommending biopsy for thyroid nodules observed at US compared with radiologists who use ACR TI-RADS.

Key Points

- For discriminating malignant and benign nodules, deep learning achieved an area under the receiver operating characteristic curve (AUC) of 0.87 (95% confidence interval [CI]: 0.76, 0.95), which is comparable to the AUC of 0.91 (95% CI: 0.82, 0.97) for a consensus of three American College of Radiologists (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) committee experts ($P = .42$) and the mean AUC of 0.82 (95% CI: 0.73, 0.90) for nine individual radiologists ($P = .38$).
- Our deep learning system achieved 52% specificity and 87% sensitivity in recommending biopsy for thyroid nodules compared with 51% specificity ($P = .91$) and 87% sensitivity ($P > .99$) from a consensus of three ACR TI-RADS committee experts.

learning continue to increase. Some of the challenges of thyroid US interpretation and reporting data systems such as ACR TI-RADS represent problems that may be solved through deep learning applications.

The aim of our study was to design a deep learning algorithm that uses thyroid US images to decide whether a thyroid nodule should undergo a biopsy. We also aimed to compare the performance of the algorithm to that of radiologists with varying expertise who adhere to ACR TI-RADS interpretation criteria.

Materials and Methods

Study Population

In this institutional review board–approved, Health Insurance Portability and Accountability Act–compliant study, we retrospectively analyzed a data set of thyroid nodules. The initial population included 1631 nodules in 1439 adult patients from a single institution who underwent diagnostic thyroid US examinations and US-guided FNA of a focal thyroid nodule between August 2006 and May 2010. It was refined by excluding 203 nodules in 172 patients who had initial nondiagnostic or indeterminate cytologic results and without subsequent cytologic or histologic diagnoses. Nodules in which images on one or both orthogonal planes were missing ($n = 15$) were also excluded. In addition, to facilitate nodule detection (based on a method that uses calipers), cases that did not contain images with proper caliper measurement marks (at least one caliper measurement on one plane and two on the other) were excluded ($n = 36$). This resulted in 1377 nodules from 1230 patients. In the final sets for the analysis, there were 1278 nodules from 1139 patients in the training set and 99 nodules from 91 patients in the test set (Fig 1). The 99 test nodules were not

used during algorithm development. They were analyzed by multiple readers in a previous study (5).

The US examinations were performed by using a variety of commercially available units (Antares and Elegra, Siemens Healthineers, Erlangen, Germany; ATL HDI 5000 and iU22, Philips, Best, the Netherlands; and Logic E9, GE Healthcare, Waukesha, Wis) equipped with 5–15-MHz linear array transducers.

Pathologic Ground Truth

FNA samples were obtained during standard clinical workflow and cytologic results were reviewed by pathology faculty at the institution (Washington University, St Louis, Mo). Determination of benignity or malignancy was made by using FNA results or, when available, surgical specimens. For FNA, five categories were used: malignant, suspicious for malignancy, indeterminate, benign, and nondiagnostic. We included nodules that were malignant or benign on the basis of initial FNA results or if a nodule underwent repeated FNA or surgical resection that subsequently provided confirmation of malignancy or benignity.

Image Annotation

All images in the training set were interpreted by one of two radiologists who were blinded to pathologic results. These two radiologists were later on the ACR TI-RADS steering committee and helped to develop ACR TI-RADS. The first reader (W.D.M.) had 22 years of experience and the second reader had 20 years of experience in thyroid imaging. By following the ACR TI-RADS lexicon, the readers assigned features for nodule composition, echogenicity, margins, and echogenic foci. For the echogenicity category, the readers classified 243 nodules as moderate to markedly hypoechoic, which was not compatible with the ACR TI-RADS lexicon. For these cases, a third reader (B.W.T., a board-eligible radiology fellow with specialty practice in thyroid imaging and 5 years of experience) reviewed the echogenicity feature and modified it by using the original assignment and additional imaging review. This reader also evaluated nodules for the shape feature. Eventually, all 1377 nodules were appropriately assigned to all five ACR TI-RADS categories.

Annotations for the five ACR TI-RADS feature categories for the test nodules were performed by 12 radiologists in December 2016, before the publication of ACR TI-RADS, with the readers blinded to the pathologic results. These interpretations were on the basis of images obtained on transverse and longitudinal planes, and video clips obtained on at least one plane displayed to the readers on standard computer monitors by using a website interface. Independent interpretations by three radiologists who were experts on the ACR TI-RADS committee, one of whom is a coauthor (F.N.T.), were combined into an expert consensus by using majority vote. These radiologists had between 26 and 34 years of posttraining experience.

Among the remaining nine readers, one reader (W.D.M.) had 22 years of experience and also interpreted the training cases. The other eight radiologists reported thyroid US in their clinical practice but had no knowledge of the management recommendations in ACR TI-RADS. This group included two academic radiologists with subspecialty training in US

and 20 and 32 years of practice experience, respectively. The six remaining radiologists from this group were from private practices with fellowship training in neuroradiology, women's imaging, and nuclear medicine, with experience ranging from 3 to 32 years.

On the basis of feature assessments for the five ACR TI-RADS categories from each reader, we first computed a total number of points per nodule and corresponding ACR TI-RADS risk levels. Then, according to ACR TI-RADS guidelines, we retrospectively decided whether a nodule would qualify for FNA and follow-up on the basis of nodule size and ACR TI-RADS risk level.

Deep Learning Algorithm

Our proposed deep learning algorithm had three main stages: nodule detection followed by prediction of malignancy and risk-level stratification. Figure 2 shows these stages and how they are connected. A complete description of all the components of the deep learning algorithm are provided in Appendix E1 (online).

For nodule detection, we first obtained a bounding box of a nodule by enclosing calipers included in every image (used in clinical practice for nodule measurement). To detect the calipers, we trained a Faster Region-based Convolutional Neural Network detection algorithm (13). After detecting the calipers on the US image, we extracted a square image with a fixed size margin of 32 pixels enclosing the corresponding nodule, resized the image to 160×160 pixels, and applied preprocessing (Appendix E1 [online]).

For classification, we trained a custom, multitask deep convolutional neural network. The tasks used for training were presence or absence of malignancy and all of the ACR TI-RADS features across the five categories (composition, echogenicity, shape, margin, and echogenic foci). The architecture of our common representation extraction network is shown in Figure 3. Source code of the model is available at the following link: <https://github.com/MaciejMazurowski/thyroid-us>.

During inference, we stratified the probability of malignancy returned by the network into risk levels referred to as deep learning risk levels (ie, DL2–DL5), modeled after the ones defined in ACR TI-RADS (ie, TR2–TR5).

Use of the deep learning risk level and a nodule's size resulted in a recommendation for FNA and follow-up. The size thresholds for FNA and follow-up recommendation were the same as in ACR TI-RADS. We used this step to choose the appropriate point on a receiver operating characteristic curve that considers nodule size and results in clinically relevant decisions.

Evaluation

We evaluated our deep learning algorithm and compared it with the performance of radiologists in two steps (Fig 4). First, we compared the performance of the algorithm to human readers for discriminating benign and malignant nodules alone by using the area under the receiver operating characteristic curve (AUC). This is the first and principal step of our algorithm and the ACR TI-RADS, and it does not involve nodule size. The AUC for the deep learning algorithm was calculated by using the likelihood of malignancy returned by model, and the AUC for radiologists used the total number of points computed with ACR TI-RADS. Then, for the second step, we evaluated the performance of the entire system in terms of sensitivity and specificity for recommendation of FNA and follow-up that in addition to the first step involves size-based thresholding. This two-step evaluation allows for isolating the predictive performance that is purely on the basis of the image from the final size-based recommendation step that aims to relate to the risk that malignant nodules

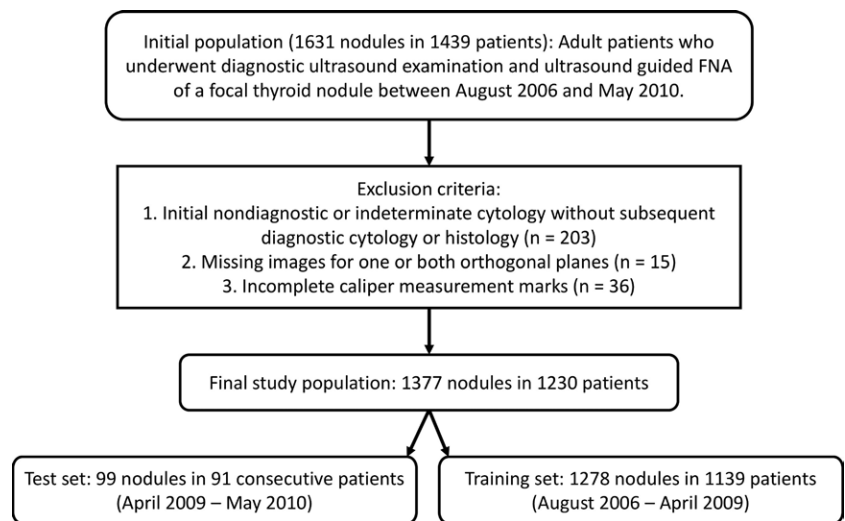


Figure 1: Flowchart of inclusion criteria for initial population and exclusion criteria for the final study population. FNA = fine-needle aspiration.

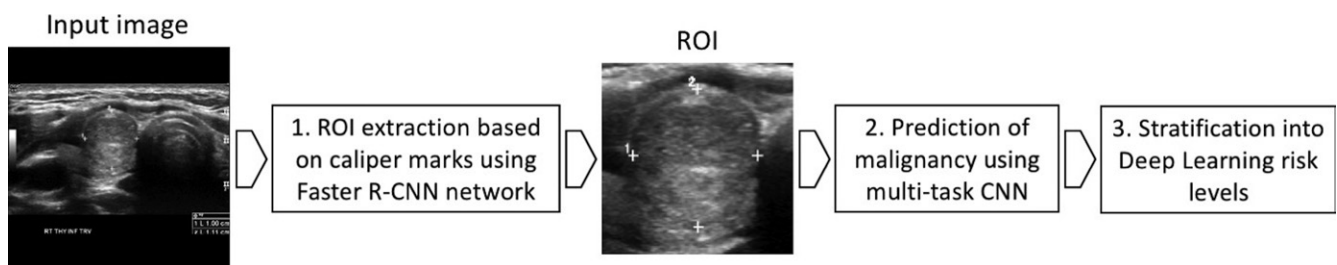


Figure 2: Flowchart of the three main processing stages of our deep learning algorithm. CNN = convolutional neural network, R-CNN = Region-based CNN, ROI = region of interest.

of different sizes pose to patients.

We performed validation of the performance of the deep learning classifier in two ways: by using a 10-fold cross-validation with our training set by pooling predictions from all 10 nonoverlapping folds and by using a hold-out test set of 99 cases. For the training set, AUC of the deep learning algorithm was compared with that of a single radiologist. On the test set, we compared the deep learning with consensus of the three ACR TI-RADS committee members and the nine other radiologists. Statistical tests for all comparisons were performed with bootstrapping.

Results

Study Population

The total number of malignant nodules was 142 (of 1377 nodules; 10.3%); there were 127 malignant nodules (of 1278 nodules; 9.9%) in the training set and 15 malignant nodules (of 99 nodules; 15%) in the test set (Table 1). The prevalence of malignant nodules between the training and test sets was not statistically significant ($P = .09$). The mean maximum nodule size for all cases was 2.6 cm (2.6 cm in the training set and 2.7 cm in the test set; $P = .53$).

Comparison of Deep Learning and Radiologists

For the training set of 1278 nodules, evaluated by using 10-fold cross-validation, the deep learning algorithm achieved an AUC of 0.78 (95% confidence interval [CI]: 0.74, 0.82) compared with 0.80 (95% CI: 0.76, 0.84; $P = .44$) for a single ACR TI-RADS committee radiologist by using ACR TI-RADS (Fig 5a).

For the test set for discriminating malignant and benign nodules, deep learning achieved an AUC of 0.87 (95% CI: 0.76, 0.95), which is comparable ($P = .42$) to that of expert consensus (0.91; 95% CI: 0.82, 0.97). The mean AUC of the nine radiologists was 0.82 (95% CI: 0.73, 0.90; not significantly lower than for deep learning, $P = .38$); the lowest AUC was 0.76 (95% CI: 0.63, 0.88) and the highest AUC was 0.85 (95% CI: 0.76, 0.94). The performance of eight of the nine individual radiologists was worse than that of deep learning; however, these differences were not statistically significant ($P > .08$). The score of each reader is provided in Table 2 and

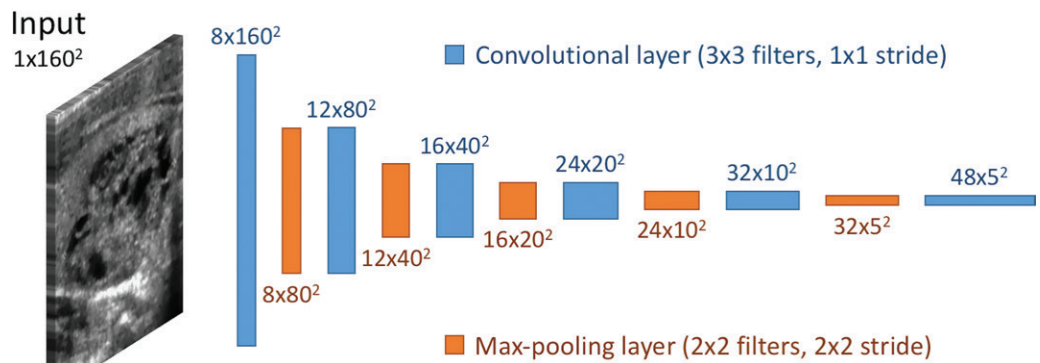


Figure 3: Convolutional neural network architecture of the network for shared representation extraction.

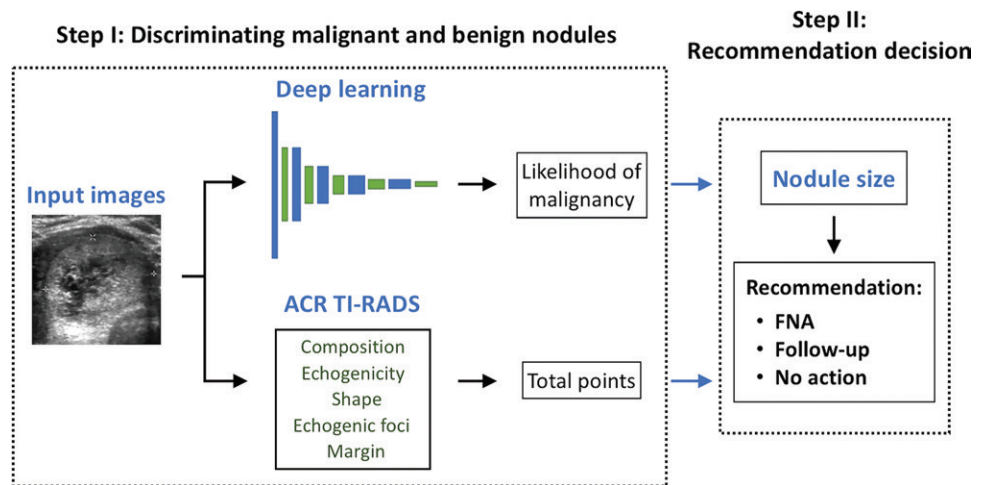


Figure 4: A diagram of the two-step decision-making process for management of thyroid nodules. ACR = American College of Radiology, FNA = fine-needle aspiration, TI-RADS = Thyroid Imaging Reporting and Data System.

the mean receiver operating characteristic curve is shown in Figure 5b.

After applying risk level stratification and size thresholds for FNA recommendation according to ACR TI-RADS, the sensitivity of the proposed deep learning algorithm was 13 of 15 (87%; 95% CI: 67%, 100%), the same as the expert consensus sensitivity of 13 of 15 (87%; 95% CI: 67%, 100%). For the nine radiologists, sensitivity ranged from 11 of 15 (73%) to 14 of 15 (93%). The differences between sensitivity of deep learning and radiologists were not statistically significant ($P > .43$). In terms of specificity, deep learning achieved 44 of 84 (52%; 95% CI: 41%, 63%), which was higher (although not significantly; $P = .91$) than expert consensus (43 of 84; 51% [95% CI: 41%, 62%]) and seven of the nine radiologists with specificity ranging from 24 of 84 (29%) to 59 of 84 (70%). The differences between specificity of deep learning and two of these seven radiologists (reader 2 and reader 8) were statistically significant ($P < .001$ and $P = .042$, respectively). The mean sensitivity and specificity for all nine radiologists was 83% (95% CI: 64%, 98%) and 48% (95% CI: 37%, 59%), respectively; both mean sensitivity and mean specificity were lower than for the deep learning algorithm (sensitivity and specificity, $P = .68$ and $.45$, respectively). Sensitivity and specificity

for FNA recommendation by all readers is provided in Table 2. Of the nodules that were misclassified by deep learning (42%; 95% CI: 33%, 53%), the nine radiologists misclassified an average of 72% (95% CI: 59%, 83%) of nodules. However, of the nodules misclassified by radiologists (average error rate, 47%; 95% CI: 37%, 56%), deep learning misclassified 66% (95% CI: 53%, 77%) of nodules. This shows a notable overlap in the misclassified cases and somewhat lower misclassification rate by the deep learning algorithm compared with that of the radiologists.

When recommending follow-up for nodules stratified into risk levels and when using size thresholds according to ACR TI-RADS, deep learning performed similarly to the radiologists. Its sensitivity was 14 of 15 (93%; 95% CI: 78%, 100%). Expert consensus did not miss any malignant nodules for recommending follow-up and achieved specificity 34 of 84 (40%; 95% CI: 30%, 51%). Similar specificity ($P = .74$) was obtained by the deep learning algorithm (specificity, 32 of 84; 38%; 95% CI: 28%, 49%). For the remaining nine readers, the mean sensitivity was 97%, whereas the mean specificity was relatively low (34%). In Table 2, we provide sensitivity and specificity for follow-up recommendation by all readers.

We split the test nodules that were positive for malignancy and negative for malignancy (ie, benign) into two subsets, easy and difficult, on the basis of the performance of human raters. Ten of 15 nodules positive for malignancy were included in the easy set on the basis of unanimous correct management decisions from all 10 readers (expert consensus and nine individual radiologists). For nodules that were negative for malignancy, 39 of 84 were also included in the easy set on the basis of at least six of 10 correct management decisions for FNA recommendation. These selections resulted in two subsets, one with 49 easy nodules (10 nodules positive for malignancy and 39 nodules negative for malignancy) and the other with 50 difficult nodules (five nodules positive for malignancy and 45 nodules negative for malignancy). Figure 6 compares the performance of deep learning and radiologists on a subset of easy (Fig 6a) and difficult (Fig 6b) test nodules. Deep learning achieved higher AUC than radiologists for the difficult nodules (0.92 vs 0.70, respectively; $P = .02$) and similar AUC for the easy nodules (0.89 vs 0.92, respectively; $P = .59$). Expert consensus and deep learning performed similarly for the difficult nodules (AUC, 0.90 [95% CI: 0.72, 1.00] vs 0.92 [95% CI: 0.80, 1.00], respectively; $P = .96$). However, for the easy nodules, the deep learning AUC (0.89; 95% CI: 0.75, 0.98) was slightly lower than for expert consensus (0.96; 95% CI: 0.89, 0.99; $P = .16$).

Discussion

Interpretation of nodules at thyroid US is time consuming and has interreader variability. In our study, we developed a deep learning algorithm to provide management recommendations for thyroid nodules observed on US images and compared its performance with radiologists who adhered to American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) guidelines. We showed that the performance

Table 1: Population Statistics according to Malignant Nodule Class

Parameter	All Nodules ($n = 1377$)	Training Nodules ($n = 1278$)	Test Nodules ($n = 99$)
Mean age of patient (y)	53.2 ± 14.0	53.2 ± 13.9	52.3 ± 14.0
Mean nodule size (cm)	2.6 ± 1.5	2.6 ± 1.5	2.7 ± 1.3
No. of malignant nodules	142 (10.3)	127 (9.9)	15 (15)

Note.—Data in parentheses are percentages; mean data are ± standard deviation.

of the algorithm was similar to that of consensus of three expert readers by achieving sensitivity of 87% (95% confidence interval [CI]: 67%, 100%) and specificity of 52% (95% CI: 41%, 63%).

The most valuable aspect of the deep learning algorithm is the ability to improve specificity of thyroid nodule biopsy recommendations. In a study that compared the recommendations of eight radiologists for 100 nodules, Hoang et al (5) found that ACR TI-RADS offered a meaningful reduction in the number of thyroid nodules recommended for biopsy and improved specificity. In our study, we show that deep learning maintains or provides improvement in specificity compared with radiologists who use ACR TI-RADS, which suggests that the proposed algorithm offers performance markedly higher than radiologists who do not use ACR TI-RADS.

Our results add to the growing body of evidence demonstrating the potential power of deep learning when applied to thyroid US. Chi et al (14) showed that a system that uses imaging features extracted with a deep convolutional neural network can achieve accuracy greater than 99% for the binary task of classifying thyroid nodules on US images to ACR TI-RADS categories 1 and 2 versus all categories. Even though the performance seems to be outstanding, it refers to a greatly simplified task of predicting proxy labels. However, our ground truth used for both the training and testing nodule subsets relied on cytologic and pathologic results. In another study, Ma et al (15) used a large data set of over 8000 thyroid nodules with malignant and benign status confirmed either by operation or FNA result. The proposed deep learning algorithm that required manual nodule segmentation resulted in high sensitivity (82%) and specificity (84%); however, nodule sizes were not considered in the evaluation. The malignancy rate was also high in that study (15) and not reflective of a typical cohort of thyroid nodules undergoing thyroid US or biopsy. However, our study compared fully autonomous decisions made by a deep learning algorithm to radiologists.

A deep learning algorithm for prediction of malignancy could make a difference in clinical practice. First, for a given image, our algorithm will always provide the same prediction. Therefore, it will eliminate a substantial interreader variability that has been observed for this task even when the ACR TI-RADS system is used. Second, the algorithm could reduce the time required for interpretation of thyroid nodules, which puts some strain on radiology departments. Finally, deep learning may perform better than some radiologists who interpret thyroid US images in clinical practice, although a larger study is needed to confirm this.

The ACR TI-RADS system consists of two steps. The first step, on the basis of specific features of the nodules, estimates the likelihood that the lesion is malignant. The second step triages

nodules for biopsy or follow-up on the basis of the likelihood estimated in the first step and nodule size. Our deep learning system replaces only the first step and uses the same size-based triaging in the second stage. Whereas this design decision was important to allow for a fair comparison of our system with ACR TI-RADS in the proper clinical setting, to some extent it limits the system to the decision-making framework of ACR TI-RADS. Future improvement that considers the interactions between tumor size and more detailed features of the nodules could provide additional gains in performance in terms of sensitivity and specificity.

Our study had limitations. Our final test set of 99 nodules (15 nodules positive for malignancy and 84 nodules negative for malignancy) as well as easy and difficult test subsets contained a small number of nodules, which resulted in wide CIs. This limitation was alleviated by a cross-validation experiment on the larger training set (127 nodules positive for malignancy and 1151 nodules negative for malignancy), which showed results that were consistent with those from the test set in terms of the comparable performance of our algorithm with the radiologist who had the highest performance. Another limitation was that we noticed some differences in performance between the test set and the training set. This was not an indication of a high-bias model (ie, underfitting) because it was the case for both deep learning and the radiologist. We believe that the main reason for this difference is that the nodules from the training set were on average more difficult to interpret, which was corroborated by additional exploration of the data including evaluation of the discriminative power of features. Whereas the overall performance of all predictors (deep learning and radiologists) differed

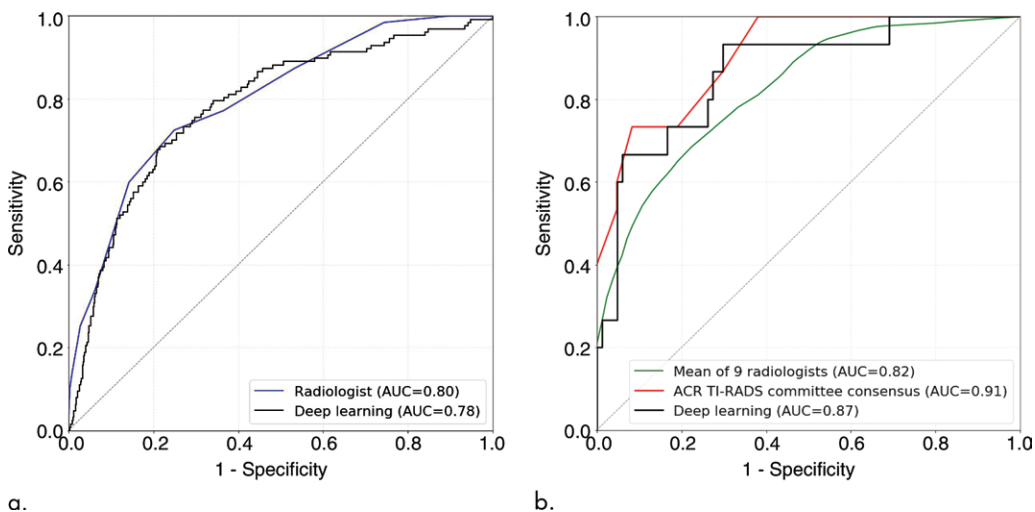


Figure 5: Areas under the receiver operating characteristic curves (AUCs) of (a) deep learning evaluated by using 10-fold cross-validation for 1278 training nodules compared with a single radiologist who used the American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) and (b) deep learning evaluated for 99 test nodules compared with expert consensus of three ACR TI-RADS committee members and nine radiologists who used ACR TI-RADS.

between the test set and the training set. This was not an indication of a high-bias model (ie, underfitting) because it was the case for both deep learning and the radiologist. We believe that the main reason for this difference is that the nodules from the training set were on average more difficult to interpret, which was corroborated by additional exploration of the data including evaluation of the discriminative power of features. Whereas the overall performance of all predictors (deep learning and radiologists) differed

Table 2: Comparison of the Deep Learning Algorithm, ACR TI-RADS Committee Expert Readers, and Radiologists

Reader	FNA		Follow-up		AUC	Experience (y)
	Sensitivity	Specificity	Sensitivity	Specificity		
Deep learning algorithm	13/15 (87) [67, 100]	44/84 (52) [42, 62]	14/15 (93) [79, 100]	32/84 (38) [28, 49]	0.87 [0.76, 0.95]	NA
ACR TI-RADS committee expert readers (n = 3)	13/15 (87)	43/84 (51)	15/15 (100)	34/84 (40)	0.91	26–32
Radiologists (n = 9)						
Reader 1	14/15 (93)	40/84 (48)	15/15 (100)	28/84 (33)	0.91	20–25
Reader 2	13/15 (87)	24/84 (29)	15/15 (100)	14/84 (17)	0.76	20
Reader 3	12/15 (80)	40/84 (48)	15/15 (100)	27/84 (32)	0.85	13
Reader 4	12/15 (80)	40/84 (48)	15/15 (100)	28/84 (33)	0.83	13
Reader 5	11/15 (73)	49/84 (57)	14/15 (93)	34/84 (40)	0.78	3
Reader 6	11/15 (73)	59/84 (70)	13/15 (87)	51/84 (61)	0.85	32
Reader 7	12/15 (80)	42/84 (50)	15/15 (100)	33/84 (39)	0.81	4
Reader 8	13/15 (87)	32/84 (38)	14/15 (93)	19/84 (23)	0.79	32
Reader 9	14/15 (93)	37/84 (44)	15/15 (100)	26/84 (31)	0.83	20
Mean values for readers 1–9 (%)	83 [64, 98]	48 [37, 59]	97 [90, 100]	34 [24, 46]	0.82 [0.73, 0.90]	17

Note.—Unless otherwise indicated, data are numerator/denominator, data in parentheses are percentages, and data in brackets are 95% confidence intervals. The readers used the test set of 99 nodules. ACR = American College of Radiology, AUC = area under the receiver operating characteristic curve, FNA = fine-needle aspiration, NA = not applicable, TI-RADS = Thyroid Imaging Reporting and Data System.

between the two sets, the relative trends between radiologists and our algorithm remained. Regarding the study population, all nodules used in our study underwent FNA because of findings suspicious for malignancy or US findings that were indeterminate, and not on the basis of ACR TI-RADS guidelines. In addition, no large-scale test set from external institutions was available for comparison and to assess for generalization to a broader population of patients and nodules.

In summary, deep learning algorithms may be promising tools in the decision-making process for assessment of thyroid nodules. More studies are needed to further validate our findings.

Acknowledgments: We thank Fernando J. Boschini, MD, Nirvikar Dahiya, MD, Jill E. Langer, MD, Justin R. Newman, MD, Carl C. Reading, MD, Daniel R. Scanga, MD, Sharlene A. Teefey, MD, Robert C. Vogler, MD, and four other radiologists who interpreted the test set of thyroid nodules as part of previously published work.

Author contributions: Guarantors of integrity of entire study, M.B., M.A.M.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, M.B., B.W.T., J.K.H., W.D.M.; clinical studies, D.T., W.D.M.; experimental studies, M.B., B.W.T., M.A.M.; statistical analysis, M.B., D.T., M.A.M.; and manuscript editing, all authors

Disclosures of Conflicts of Interest: M.B. disclosed no relevant relationships. B.W.T. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed an invention disclosure to the Duke University Office of Technology and Licensing. Other relationships: disclosed no relevant relationships. J.K.H. disclosed no relevant relationships. D.T. disclosed no relevant relationships. F.N.T. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed money paid to author for expert testimony from Starnes, Davis, Florie; disclosed speaking honoraria from the American College of Radiology. Other relationships: disclosed no relevant relationships. W.D.M. disclosed no relevant relationships. M.A.M. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed an invention disclosure to the Duke University Office of Technology and Licensing as well as advising relationship with Gradient Health. Other relationships: disclosed no relevant relationships.

References

- 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):1–133.
- Hoang JK, Middleton WD, Farjat AE, et al. Interobserver Variability of Sonographic Features Used in the American College of Radiology Thyroid Imaging Reporting and Data System. *AJR Am J Roentgenol* 2018;211(1):162–167.
- Vaccarella S, Franceschi S, Bray F, Wild CP, Plummer M, Dal Maso L. Worldwide Thyroid-Cancer Epidemic? The Increasing Impact of Overdiagnosis. *N Engl J Med* 2016;375(7):614–617.
- 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(5):587–595.

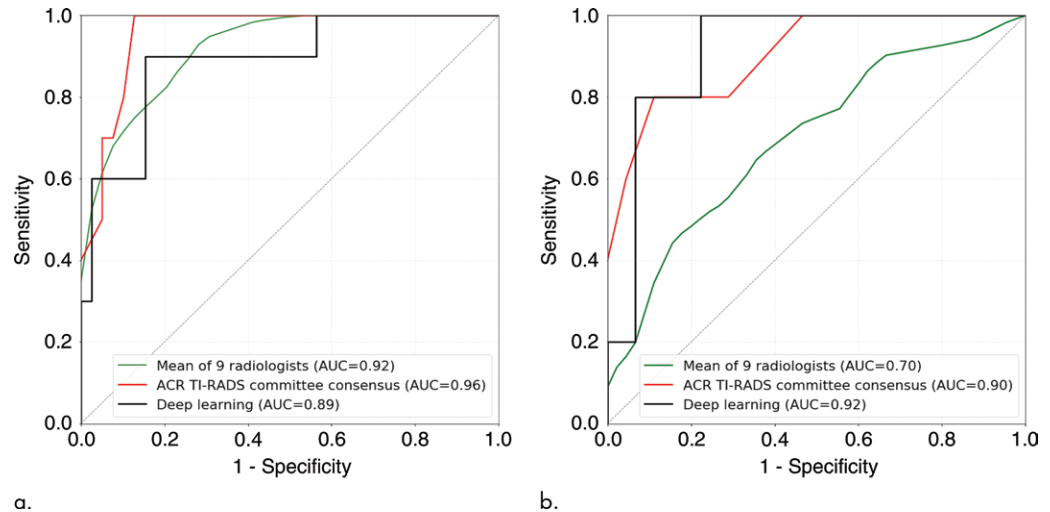


Figure 6: Areas under the receiver operating characteristic curves (AUCs) comparing deep learning, American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) committee consensus, and radiologists for (a) 49 easy test nodules and (b) 50 difficult test nodules. The easy nodules (a) include 10 malignant nodules on the basis of unanimous correct management decisions from nine readers and expert consensus and 39 benign nodules on the basis of at least six of 10 correct management decisions for fine-needle aspiration recommendation. The difficult nodules (b) include the remaining five malignant and 45 benign test nodules.

- Hoang JK, Middleton WD, Farjat AE, et al. Reduction in thyroid nodule biopsies and improved accuracy with American college of radiology thyroid imaging reporting and data system. *Radiology* 2018;287(1):185–193.
- Griffin AS, Mitsky J, Rawal U, Bronner AJ, Tessler FN, Hoang JK. Improved Quality of Thyroid Ultrasound Reports After Implementation of the ACR Thyroid Imaging Reporting and Data System Nodule Lexicon and Risk Stratification System. *J Am Coll Radiol* 2018;15(5):743–748.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521(7553):436–444.
- Esteve A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542(7639):115–118 [Published correction appears in *Nature* 2017;546(7660):686].
- 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(22):2402–2410.
- Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. *RadioGraphics* 2017;37(2):505–515.
- Lee H, Tajmir S, Lee J, et al. Fully automated deep learning system for bone age assessment. *J Digit Imaging* 2017;30(4):427–441.
- Mazurowski MA, Buda M, Saha A, Bashir MR. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *J Magn Reson Imaging* 2019;49(4):939–954.
- Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv Neural Inf Process Syst* 2015; 91–99. <https://dl.acm.org/citation.cfm?id=2969250>.
- Chi J, Walia E, Babyn P, Wang J, Groot G, Eramian M. Thyroid Nodule Classification in Ultrasound Images by Fine-Tuning Deep Convolutional Neural Network. *J Digit Imaging* 2017;30(4):477–486.
- Ma J, Wu F, Zhu J, Xu D, Kong D. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. *Ultrasonics* 2017;73:221–230.
- He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proc IEEE Conf Comput Vis Pattern Recognit*, 2016, 770–778.
- Lin TY, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context. *Eur Conf Comput Vis*, 2014; 740–755.
- Buades A, Coll B, Morel JM. A non-local algorithm for image denoising. *Comput Vis Pattern Recognition*, 2005 CVPR 2005 IEEE Comput Soc Conf, 2005; 60–65.
- Coupé P, Hellier P, Kervrann C, Barillot C. Nonlocal means-based speckle filtering for ultrasound images. *IEEE Trans Image Process* 2009;18(10):2221–2229.
- Caruana R. Multitask learning. In: Thrun S, Pratt L, eds. *Learning to Learn*. Boston, Mass: Springer, 1998; 95–133.
- Nitish S, Hinton GE, Alex K, Ilya S Sr. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;15(1):1929–1958. <http://jmlr.org/papers/v15/srivastava14a.html>.
- Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. *arXiv Prepr arXiv:1708.02002*. <https://arxiv.org/abs/1708.02002>. Published August 7, 2017. Accessed DATE.
- Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw* 2018;106:249–259.