Thyroid Nodule Malignancy Risk Stratification Using a Convolutional Neural Network

Matthew T. Stib, MD, Ian Pan, MA,* Derek Merck, PhD,* William D. Middleton, MD,† and Michael D. Beland, MD**

Downloaded from http://journals.lww.com/ultrasound-quarterly by BhDMf5ePHKbH4TTImqenVBZZxeh5YHRLnBGwiyl2rGSxcaJ227VZwXA8w9dlf3e on 07/12/2020

Abstract: This study evaluates the performance of convolutional neural networks (CNNs) in risk stratifying the malignant potential of thyroid nodules alongside traditional methods such as American College of Radiology Thyroid Imaging Reporting and Data System (ACR TIRADS). The data set consisted of 651 pathology-proven thyroid nodules (500 benign, 151 malignant) from 571 patients collected at a single tertiary academic medical center. Each thyroid nodule consisted of two orthogonal views (sagittal and transverse) for a total of 1,302 grayscale images. A CNN classifier was developed to identify malignancy versus benign thyroid nodules, and a nested double cross validation scheme was applied to allow for both model parameter selection and for model accuracy evaluation. All thyroid nodules were classified according to ACR TIRADS criteria and were compared with their respective CNN-generated malignancy scores. The best performing model was the MobileNet CNN ensemble with an area under the curve of 0.86 (95% confidence interval, 0.83–0.90). Thyroid nodules within the highest and lowest CNN risk strata had malignancy rates of 81.4% and 5.9%, respectively. The rate of malignancy for ACR TIRADS ranged from 0% for TR1 nodules to 60% for TR5 nodules. Convolutional neural network malignancy scores correlated well with TIRADS levels, as malignancy scores ranged from 0.194 for TR1 nodules and 0.519 for TR5 nodules. Convolutional neural networks can be trained to generate accurate malignancy risk scores for thyroid nodules. These predictive models can aid in risk stratifying thyroid nodules alongside traditional professional guidelines such as TIRADS and can function as an adjunct tool for the radiologist when identifying those patients requiring further histopathologic workup.

Key Words: artificial intelligence, machine learning, thyroid nodules, ultrasound

Abbreviations: ACR TIRADS = American College of Radiology Thyroid Imaging Reporting and Data System, AUC = area under the curve, CNN = convolutional neural network, FNA = fine-needle aspiration

(*Ultrasound Quarterly* 2020;36: 164–172)

Received for publication October 28, 2019; accepted March 17, 2020.

*Department of Diagnostic Imaging, Rhode Island Hospital, Alpert Medical School of Brown University, Providence, RI; and †Mallinckrodt Institute of Radiology, School of Medicine, Washington University, St. Louis, MO.

The authors declare no conflicts of interest.

Address correspondence to: Matthew T. Stib, MD, Department of Diagnostic Imaging, Rhode Island Hospital, Alpert Medical School of Brown University, 593 Eddy St, Providence, RI 02903 (e-mail: matthew.stib@lifespan.org).

Copyright © 2020 Wolters Kluwer Health, Inc. All rights reserved.

DOI: 10.1097/RUQ.0000000000000501

The detection threshold for thyroid nodules has decreased in the age of diagnostic imaging, primarily through the improved technology of high-resolution thyroid ultrasound. This observation correlates with reports that up to 67% of asymptomatic subjects were found to have thyroid nodules at ultrasound and 50% at autopsy, compared with a 4% detection rate with palpation alone.[1]

Despite this high prevalence, studies have found that 88% to 98% of incidentally detected thyroid nodules are benign.[2,3] In addition, many thyroid cancers may not be clinically relevant, as demonstrated by autopsy studies showing clinically occult thyroid carcinoma in up to 5% of patients.[4]

Several professional organizations have published guidelines to standardize vocabulary and risk stratification based on various ultrasound characteristics. The American College of Radiology Thyroid Imaging Reporting and Data System (ACR TIRADS) is one widely used system that stratifies the malignancy risk of thyroid nodules into five different categories (TR1–TR5, from least to most suspicious) based on various sonographic imaging features (composition, echogenicity, margin, echogenic foci, and shape).[5] However, this system remains imperfect because of subjectivity and reliance of interobserver variability in interpretation, leading to unnecessary fine-needle aspiration (FNA) or surgery for definitive diagnosis.[6] As a result of the subjectivity in guideline interpretations, there has been increasing interest in computer-aided diagnosis for thyroid nodule malignancy prediction by taking advantage of objective sonographic features.[7]

Recent advances in deep learning have propelled the technology to the forefront of image classification in virtually every domain, driven largely by the use of convolutional neural networks (CNNs) in the ImageNet Large Scale Visual Recognition Competition.[8] These models are powerful in that they process the raw image directly and learn hierarchical imaging features specific to the given classification task, without the need for manual feature extraction. Convolutional neural networks and their variants have set standards for state-of-the-art performance in various image-related tasks, such as classification, segmentation, and object detection.[9,10]

The success of CNNs in other domains has inspired their adoption in radiology artificial intelligence. There have been a number of promising results to date, including automated detection of various findings in chest radiographs,[11–13] identification of acute findings in head computed tomography,[14] bone age assessment in pediatric hand radiographs,[15,16] and brain tumor segmentation in magnetic resonance imaging.[17] In addition,

Copyright © 2020 Wolters Kluwer Health, Inc. All rights reserved.

successive generation of CNNs has shown smaller performance gains on image classification benchmarks, leading to a new focus on creating smaller, lightweight CNNs (MobileNet), which have shown comparable performance and could be used in real-time medical applications.[18,19]

This study aims to evaluate the performance of a CNN in risk stratifying the malignancy potential thyroid nodules alongside traditional methods such as ACR TIRADS.

MATERIALS AND METHODS

Data

This was a retrospective study of a consecutively acquired data set of patients who underwent thyroid sonography and FNA and who agreed to participate in the study at a single tertiary academic institution, with details previously described.[20] The sonographic images were deidentified and compliant with the Health Insurance Portability and Accountability Act. The grayscale sonographic images were obtained using a high-frequency linear transducer on a variety of commercially available machines including Siemens Healthineers (Malvern, PA), General Electric (Chicago, IL) and Phillips (Netherlands). During acquisition, specific attention was paid to obtaining optimal imaging for nodule characteristics (e.g., composition, echogenicity, margin types, echogenic foci). Although image acquisition was obtained heterogeneously using different ultrasound machine vendors and a variety of sonographers likely using slightly different techniques, this variability likely decreased the risk of model overfitting and increased generalizability.[21] Each thyroid nodule consisted of at least one sagittal and one transverse view. The entire data set used in this study consisted of 651 nodules (500 benign/151 malignant) from 571 patients totaling 1,302 grayscale images acquired between August 2006 and May 2010. Nodules measured 4 to 98 mm in maximum diameter with the majority of nodules (93%) measuring at least 1.0 cm in the largest dimension. The mean patient age was 52.9 years (SD, 14.2 years) and 59% were women.

Each thyroid nodule used in this study had a correlate ground-truth label based on definitive cytological results (malignant/benign) from FNA and/or surgical pathology following thyroidectomy. All of the malignant nodules had surgical pathology. There were 151 malignant nodules, the majority consisted of papillary carcinoma (n = 135, 89.4%), followed by follicular (n = 6, 4%) and other (n = 10, 6.6%). Decisions to biopsy a particular nodule were clinically determined on an individual patient basis following the application of the Society of Radiologists in Ultrasound guidelines.[22]

Each thyroid nodule in the data set was independently analyzed retrospectively for sonographic characteristics using the ACR lexicon[5] and given an associated ACR TIRADS level (TR1–TR5) as described in the previous study.[20]

Preprocessing

For each known pathologically proven thyroid nodule, a corresponding sagittal and transverse grayscale sonographic image was manually selected from the full stack of images from the original clinical thyroid ultrasound examination. The nodule and immediately adjacent surrounding tissue were manually cropped out from the image as a method to increase the signal to noise from the surrounding structures and converted to a JPEG image (Fig. 1). These images were then resized to 224 × 224 pixels, and pixel values were normalized to a range between −1 and 1.

Model Architecture

We compared the MobileNet and ResNet50 CNN architectures in this study. MobileNets are efficient CNN architectures

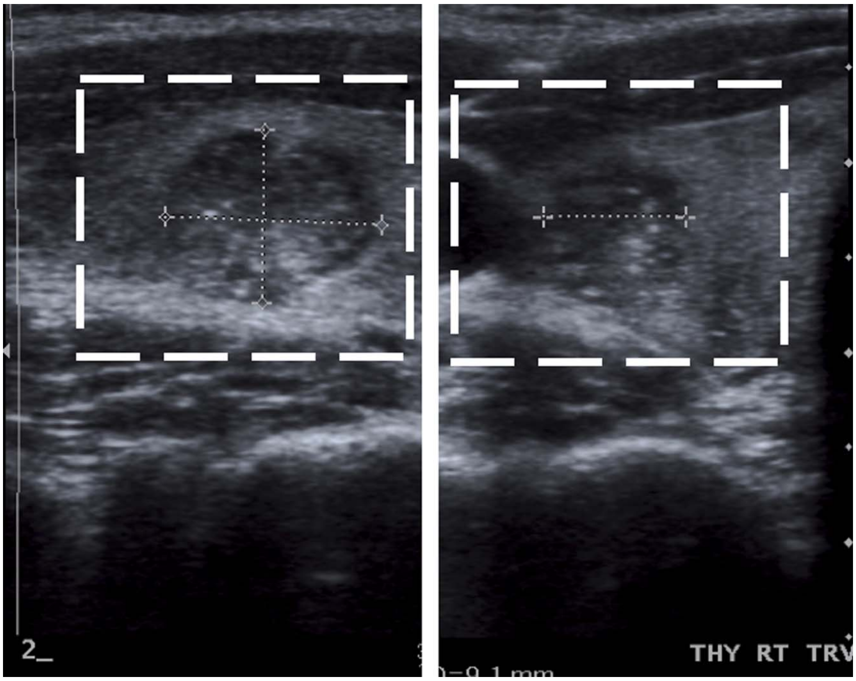

FIGURE 1. A 62-year-old woman with thyroid nodule. Manual region of interest cropping (dotted lines) was performed to focus the network on areas of interest (left, sagittal view; right, transverse view).

© 2020 Wolters Kluwer Health, Inc. All rights reserved.

Copyright © 2020 Wolters Kluwer Health, Inc. All rights reserved.

intended for embedded mobile applications that have considerably fewer parameters, that is, weights and biases that a model learns from training data, than other commonly used "deeper" CNN architectures such as Inception-V3, ResNet50, and DenseNet.[23] Convolutional neural networks were instantiated with pretrained parameters obtained by first training the network on 1.2 million color images of common objects from ImageNet (http://www.image-net.org). These parameters were then modified by training the CNN on the current task of thyroid nodule malignancy prediction, a process known as transfer learning.[24]

Model Training and Evaluation

Models were trained on a NVIDIA GTX 1080 Ti GPU (Santa Clara, CA) using Keras version 2.0.9 (Chollet, Francois and others, 2015, http://keras.io) and Python version 2.7 (Python Software Foundation, https://www.python.org).

A nested cross-validation scheme was used for model hyperparameter selection and generation of a malignancy score for each thyroid nodule.[25] This method consisted of two separate procedures, whereby an inner cross-validation loop was used for parameter tuning followed by an outer loop where each nodule in the data set was used once for model testing (Fig. 2). The inner loop was created by dividing the data into 10 disjoint stratified folds and with 3 separate training/validation splits following a 90%/10% distribution after reserving a 10% for a held-out test set. Each training/validation split produced a separate model, which was then used to produce a final prediction on the reserved nodules in the held-out test set. The outer loop consisted of this process being repeated 10 times so that each nodule in the data set was part of a held-out test set separate from training/validation and had its own prediction score.

Data augmentation and dropout were used to prevent overfitting.[26] Data augmentation was performed in real time, with random zooms, rotations, vertical and horizontal flips, Gaussian smoothing, and contrast adjustments. Dropout was applied after the final fully connected layer.

Optimization was performed for 3 hyperparameters (learning rate, data augmentation probability, and dropout probability) using a randomized search over 60 iterations. For each iteration, hyperparameter values were sampled from prespecified uniform distributions. Models were trained for 20 epochs using the Adam optimizer[27] with default parameters. As imbalanced classes (e.g., substantially higher percentage of benign nodules) can negatively affect model training, benign nodules were randomly sampled from the overall benign population to match the number of malignant nodules during each epoch. Models were validated on the validation fold after every epoch, and the best performing model, as measured by the area under the curve (AUC), across all epochs and hyperparameter iterations was selected for evaluation.

To obtain a prediction for a given thyroid nodule, the sagittal and transverse views were treated as separate inputs to the CNN, resulting in two predicted malignancy scores per nodule that were then averaged. The final model was an ensemble of three CNNs that averaged across each model's prediction.

Statistical Analysis

All statistical analyses were performed using R 3.4.2 (R Foundation for Statistical Computing, https://www.R-project.org). The receiver operating characteristic curve, cross-validated AUC, and its 95% confidence interval (CI) were calculated using the cvAUC R package.[28] Sensitivities and specificities for the CNN malignancy rates were calculated at five predetermined thresholds (0.1, 0.2, 0.5, 0.8, and 0.9), and the bootstrap method was used to calculate 95% CIs. Statistical significance was set to $P < 0.05$. To estimate the sensitivity and specificity of the TIRADS classification system, true-positive tests included TR4 and TR5 nodules that were histopathologically malignant, and true-negative tests were TR1 to TR3 that were histopathologically benign.

RESULTS

The results comparing the two CNNs are summarized in Table 1. MobileNet consistently outperformed ResNet50 in each experimental configuration, although improvements were not statistically significant. The MobileNet and ResNet50 two-view ensembles (MobileNet-TVE, ResNet50-TVE) achieved

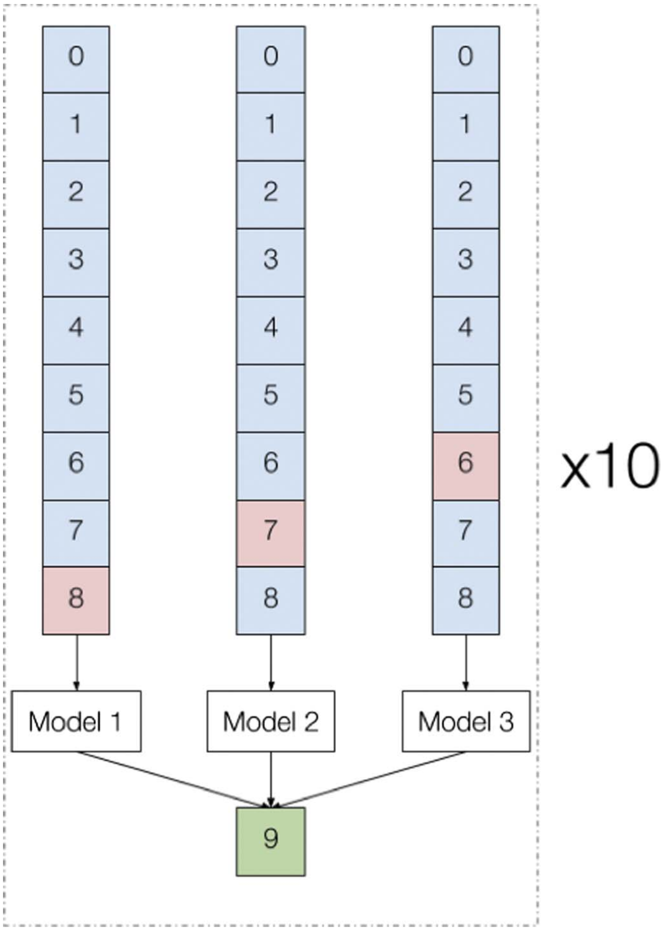

FIGURE 2. Double cross-validation scheme. Ten percent of the data are reserved for model evaluation (green). The remaining 90% is divided into training (blue) and validation (red), where the validation split is used to select the highest performing model for evaluation. This is repeated such that all folds 0 to 9 are used for evaluation, resulting in an out-of-sample prediction for each observation in the data set.

© 2020 Wolters Kluwer Health, Inc. All rights reserved.

Copyright © 2020 Wolters Kluwer Health, Inc. All rights reserved.

TABLE 1. Comparison of AUCs Between Architectures and Among Different Experimental Configurations

	Sagittal View	Transverse View	Two-View Average	Two-View Ensemble
MobileNet	0.806	0.761	0.813	0.863
Parameters: 3,229,313*	(0.766, 0.846)	(0.718, 0.806)	(0.774, 0.853)	(0.827, 0.898)
ResNet50	0.756	0.761	0.799	0.838
Parameters: 23,583,489	(0.716, 0.797)	(0.720, 0.803)	(0.762, 0.838)	(0.800, 0.877)

The differences between models were not statistically significant. Parameters refer to how many "weights" were in each neural network.

*Significantly less parameters with MobileNet versus ResNet50.

the highest performance overall, with cross-validated AUCs of 0.86 (95% CI, 0.83–0.90) and 0.84 (95% CI, 0.80–0.88), respectively ($P = 0.65$).

There was no significant model performance difference between the single transverse versus single sagittal view for either CNN. Two-view (sagittal and transverse) averaging and three-model ensembling improved performance as compared with the one-view single models for both CNNs (Table 1). A comparison of ROC curves across nine folds for MobileNet-TVE as well as a comparison between the two model architectures is illustrated in Figure 3. Over 10 folds, the test AUC ranged from 0.79 to 0.96, compared with an overall mean of 0.86. A single test fold can significantly overestimate or underestimate model performance with small data sets.[29] The median malignancy scores for benign and malignant nodules were 0.16 and 0.62, respectively, for MobileNet-TVE. A distribution of the malignancy scores is shown in Figure 4.

Sensitivities and specificities across five predetermined malignancy thresholds (0.1, 0.2, 0.5, 0.8, and 0.9) are provided in Table 2 for MobileNet-TVE and ResNet50-TVE. Table 3 illustrates the malignancy rate in five MobileNet-TVE malignancy score strata, as well as the relative risk compared with the data set's overall malignancy rate of 23.2% (151 of 651 nodules). Thyroid nodules in the highest risk stratum (malignancy score of 0.8–1.0) has a malignancy rate of 81.4% as compared with 5.94% in the lowest risk stratum (malignancy score of 0–0.2). There was not a significant correlation between nodule size and predicted malignancy score (Pearson's correlation, $r = -0.28$; $P = 0.17$).

Inference times for two-view ensembles were evaluated with (NVIDIA Tesla P100) and without (Intel Xeon CPU E5-2670) graphics processing unit (GPU) acceleration. With GPU acceleration, average inference times for 1,000 thyroid nodules for MobileNet-TVE and ResNet50-TVE were 8.9 seconds and 20.5 seconds, respectively. Without GPU acceleration, average inference times for each model increased to 86 seconds and 248 seconds. In other words, the light-weight artificial intelligence framework was three to four times faster with inference compared with the traditional artificial intelligence framework running on the same hardware.

The percentage of thyroid nodules in each TIRADS level was 11.7%, 12.6%, 19.0%, 36.4%, and 20.3%, respectively, for TR1 to TR5, and the malignancy rate in each TIRADS increased from 0% for TR1 to 60.0% for TR5. The malignancy scores predicted by the MobileNet CNN model were positively correlated with the TIRADS levels, with an average score of 0.194 for TR1 and 0.519 for TR5. The estimated sensitivity and specificity for the TIRADS classification were 97.7% and 51.8%, respectively (Table 4). We explored several cases of thyroid nodules that were moderate suspicious using TIRADS (TR4) and applied our MobileNet CNN scores to demonstrate how the combined risk stratification systems could be used for clinical application (Fig. 5).

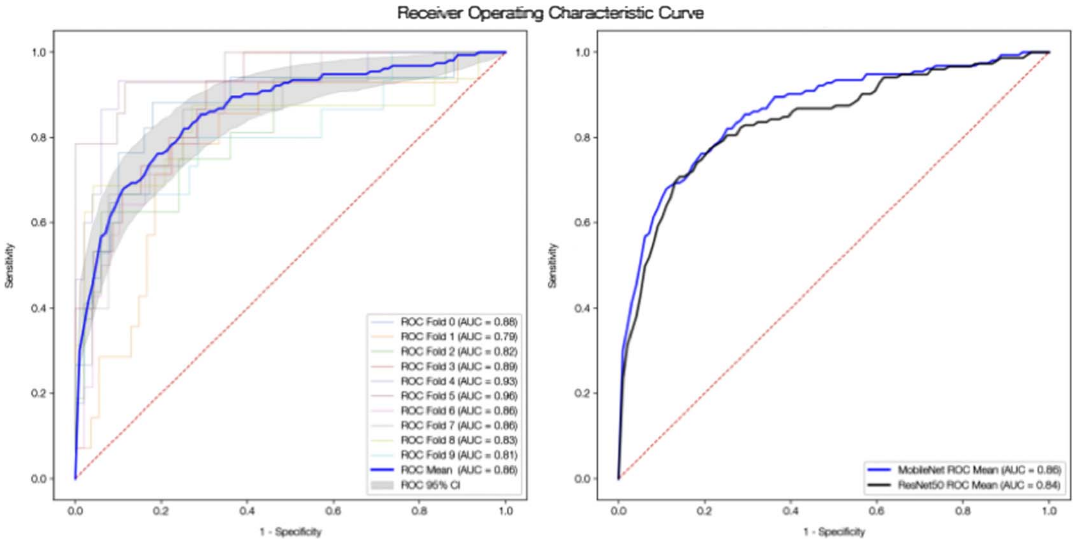

FIGURE 3. Cross-validated ROC curves. Left, MobileNet ROC curves, mean (blue; AUC, 0.86) and per fold (various colors; AUC range, 0.79–0.96). Right, MobileNet mean ROC curve (blue; AUC, 0.86) versus ResNet50 mean ROC curve (black; AUC, 0.84).

© 2020 Wolters Kluwer Health, Inc. All rights reserved.

Copyright © 2020 Wolters Kluwer Health, Inc. All rights reserved.

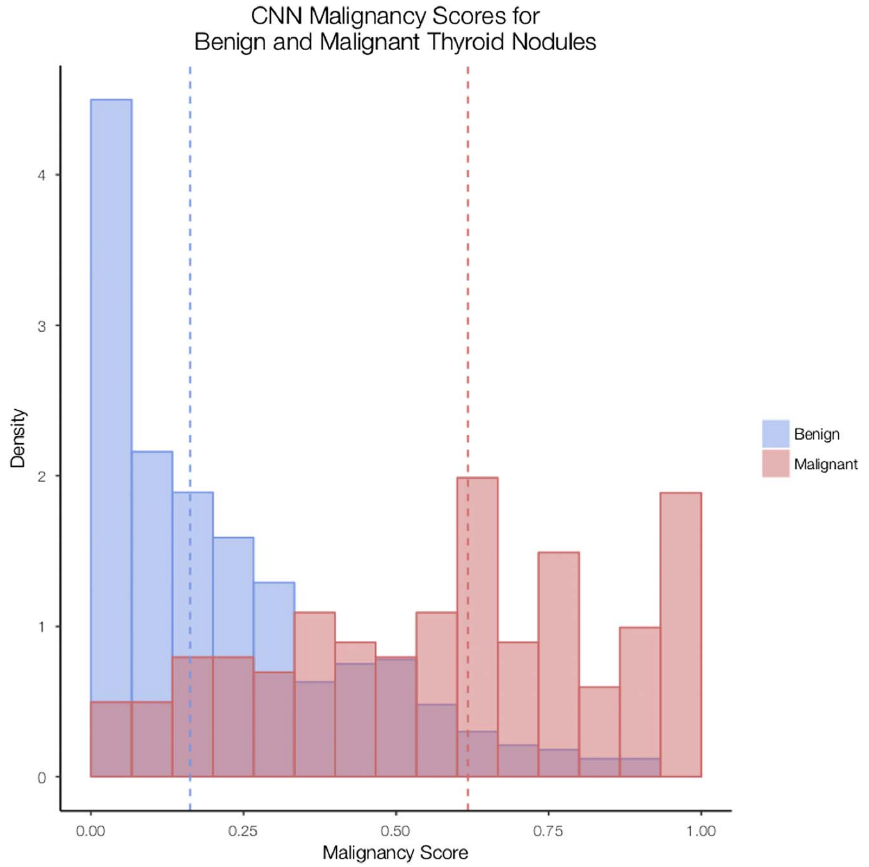

FIGURE 4. Distribution of malignancy scores for benign and malignant nodules. Dotted lines, median malignancy score.

We performed a retrospective review of the false-positive and false-negative cases and provided several examples of observed trends in Figure 6. The nodules in Figure 6A and B were examples of false-negative results with CNN malignancy scores of 0.025 and 0.014, respectively. The nodules in Figure 6C and D were examples of false-positive results with CNN malignancy scores of 0.809 and 0.886, respectively.

DISCUSSION

We explored the ability of CNNs trained on pathology-proven sonographic images to differentiate between benign and malignant thyroid nodules. Given the increasing detection rate of thyroid nodules, developing a noninvasive classification tool is important in reducing the number of unnecessary invasive diagnostic procedures. Both of the CNNs that we trained demonstrated high performance in differentiating between pathologically proved benign and malignant thyroid nodules, with the lightweight MobileNet CNN achieving an overall AUC of 0.86. The diagnostic performance of the CNN could be further optimized by adjusting the predetermined threshold score to either increase sensitivity (95.4% at a threshold score of 0.1) or specificity (99.8% at threshold score of 0.9).

Reducing the number of unnecessary FNAs is of particular clinical significance given the occult nature of many thyroid cancers,[2–4] and various professional societies have proposed guidelines attempting to optimize the tradeoff between malignant nodule detection and reducing unnecessary FNAs.[22,30–33] These guidelines rely on radiologists to qualitatively identify suspicious sonographic features, which is time-consuming and has inherent interobserver and intraobserver variability. Several recent studies have examined the diagnostic performance of

TABLE 2. Sensitivities and Specificities at Predetermined Malignancy Score Thresholds

	MobileNet-TVE		ResNet50-TVE	
Threshold Score	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)
0.1	36.3 (31.9–40.6)	95.4 (91.7–98.5)	20.7 (17.4–23.9)	95.2 (91.7–98.3)
0.2	57.3 (53.5–61.4)	87.9 (82.4–92.7)	37.6 (33.9–41.2)	93.4 (89.5–97.0)
0.5	88.3 (85.4–91.1)	61.7 (54.2–69.1)	78.8 (75.0–82.1)	71.6 (64.9–77.8)
0.8	98.4 (97.4–99.4)	23.3 (17.2–29.6)	98.2 (97.0–99.2)	25.9 (20.0–32.0)
0.9	99.8 (99.3–100)	16.0 (10.5–21.9)	99.6 (99.0–100)	12.0 (7.51–16.6)

© 2020 Wolters Kluwer Health, Inc. All rights reserved.

Copyright © 2020 Wolters Kluwer Health, Inc. All rights reserved.

TABLE 3. Malignancy Rates in Five Malignancy Score Strata (MobileNet CNN)

Malignancy score	0–0.2	0.2–0.4	0.4–0.6	0.6–0.8	0.8–1.0
Malignancy rate*	5.94% (0.26)	18.2% (0.78)	29.5% (1.27)	65.7% (2.83)	81.4% (3.51)

*Malignancy rate indicates risk relative to overall malignancy rate for all nodules.

these professional guidelines[20,34] in detecting thyroid malignancy. Ha et al.[35] found that the ACR TIRADS guidelines resulted in the lowest rate of unnecessary thyroid FNAs with a false-positive rate of 32.7% and a false-negative rate of 25.3%. In comparison, the diagnostic performance of our MobileNet CNN model set at a malignancy threshold score of 0.5 would result in a much lower false-positive biopsy rate of 11.7% without much sacrifice in the false-negative rate, only 13 percentage points higher at 38.3%. Alternatively, increasing the malignancy score cutoff for FNA even higher could vastly decrease the false-positive rate (1.6% at a threshold of 0.8) but with a higher false-negative rate. However, depending on the clinical situation, this tradeoff favoring a very low false-positive rate might be preferred given the indolent nature of most thyroid cancers, as nodules can safely and noninvasively be serially monitored via sonography for increased signs of malignancy.[36]

Applying an objective CNN-generated malignancy score in combination with traditional classification systems such as the TIRADS guidelines has the potential to further improve risk stratification of thyroid nodules as compared with each method alone. The ACR TIRADS schema applied to the nodules in this study resulted in an overall sensitivity of 97.7 with a significant false-positive rate of 48.2%. Convolutional neural network–generated malignancy scores could assist a traditional professional guideline system such as TIRAD on whether or not to proceed to FNA as an added tool, as depicted in Figure 5. This is an example of two "moderately suspicious" TR4 nodules, which were both biopsied, but only nodule B was found to be malignant. By also examining the CNN malignancy scores of 0.018 and 0.68 for nodules A and B, respectively, one could use the very low CNN score for nodule A to bolster a recommendation to forgo a biopsy despite its large 40 mm size. Conversely, one could more definitely proceed with biopsy for the smaller 6 mm nodule B given its high CNN score.

Other studies have also explored the use of computer-aided diagnosis in determining thyroid nodule malignancy. Liu et al.[37] designed a Bayesian classifier that combined demographics and sonographic features to predict malignancy probabilities of thyroid nodules, achieving a similar AUC of 0.85 on 93 nodules. Wu et al.[38] adopted a machine learning approach using a Bayesian classifier and support vector machine using manually extracted sonographic features as inputs to classify thyroid nodules, achieving AUCs of up to 0.91 on 970 patients. However, both of these approaches required experienced radiologist to manually identify and label relevant sonographic features for model inputs. Chi et al.[39] used deep learning to discriminate between TIRADS levels 1 and 2 versus TIRADS levels 3 to 5, achieving an AUC of 0.99. This study resulted in nearly perfect TIRADS categorical discrimination, but notably, the TIRADS scores themselves were not correlated with pathologic ground truth.

We reviewed the misclassifications made by our models and found the following observations demonstrated in Figure 6. A common trend among the false-negative nodules with the lowest CNN-predicted malignancy scores was that they contained at least a partial cystic composition. This highlights an interesting point in that the model seems to correctly identify that the cystic composition is associated with benignity, however, at the expense of underestimating the malignant potential of the adjacent solid components. Common trends regarding several of the false positive nodules with the highest CNN-predicted malignancy scores were that these nodules contained calcifications and had hypoechogenic echotextures. Although these particular nodules were pathologically benign, it seems that the models were selecting appropriate features associated with malignancy, which was also reflected by their corresponding moderately suspicious TIRADS scores.

There are several limitations to our study including general machine learning aspects as well as related to the specific problem of thyroid nodule malignancy prediction. First, this study was trained and validated on a data set from a single institution with a relatively limited data set for the field of deep learning. Even though data were collected on a variety of different ultrasound machines with a diverse set of human sonographers performing the scans, generalizability could be improved by validating our models on additions sources of data and we are planning further investigation using unseen data from a separate institution. Also, our model's 2-dimesional input likely does not

TABLE 4. Application of ACR TIRADS Classification Schema

TIRADS Level	Number	Total, %	Malignant, %	Average CNN Malignancy Score (SD)
TR1	76	11.7	0.0	0.194 (0.19)
TR2	82	12.6	3.7	0.160 (0.17)
TR3	124	19.0	14.5	0.250 (0.24)
TR4	237	36.4	21.1	0.293 (0.26)
TR5	132	20.3	60.0	0.519 (0.29)
Overall TIRADS sensitivity:	97.7%			
Overall TIRADS specificity:	51.8%			

SD, standard deviation.

© 2020 Wolters Kluwer Health, Inc. All rights reserved.

Copyright © 2020 Wolters Kluwer Health, Inc. All rights reserved.

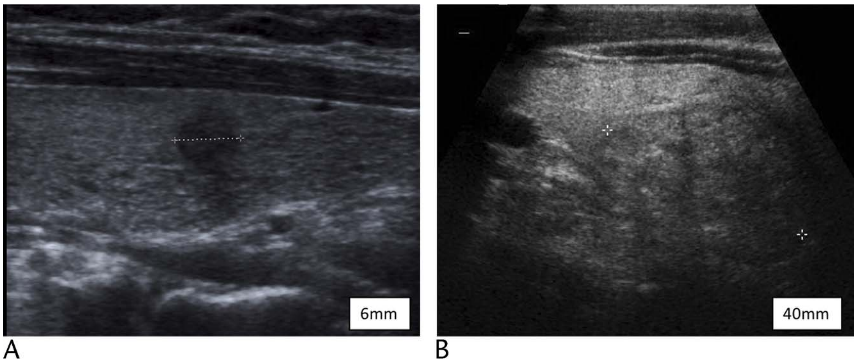

FIGURE 5. Two patients with moderately suspicious TIRADS category 4 thyroid nodules, pathology-proven malignant (*A*) and benign (*B*). (*A*) A 71-year-old woman with TR4 thyroid nodule. Sagittal view of the left thyroid gland demonstrating a 6 mm nodule, TIRADS 4 (five points, very hypoechoic [3] and solid [2]), CNN malignancy score of 0.680, and pathology-proven malignant. (*B*) A 68-year-old man with TR4 thyroid nodule. Sagittal view of the right thyroid gland demonstrating a 40 mm nodule, TIRADS 4 (five points, hypoechoic [2], solid [2], and macro calcifications [1]), CNN malignancy score of 0.018, and pathology-proven benign.

capture the complete 3-dimensional heterogeneity of a thyroid nodule's composition. Furthermore, most machine learning applications in medicine are the opaque "black-box" results of model predictions, which make it hard to explain which features the models found most relevant, thus, decreasing the generalizability to new data without further verification methods. Another limitation is that our models are not completely automated because they require the preprocessing step of manual identification and cropping of nodules from the full sonographic image. Using this CNN in clinical workflow would require drawing a region of interest around each thyroid nodule before entering the model as input data. However, this step would not be a significant additional burden since a similar workflow already is performed in clinical practice when a technologist or radiologist manually selecting suspicious nodules from a scan and performs maximal diameter measurements of those selected nodules. A bounding box could

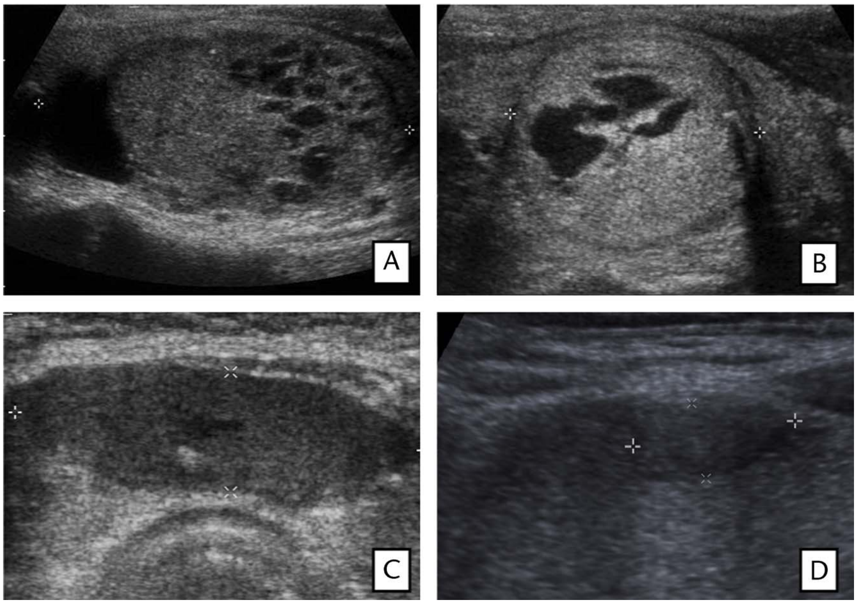

FIGURE 6. Two patients (*A* and *B*) with false-negative predictions of thyroid nodules by the CNN (pathologically malignant) and two patients (*C* and *D*) with false-positive predictions of thyroid nodules by the CNN (pathologically benign). A, A 56-year-old woman with false-negative CNN prediction. Sagittal view of the right thyroid gland demonstrating a mixed cystic and solid nodule, CNN malignancy score of 0.025, and TIRADS 3 (three points, hypoechoic [2] and mixed cystic/solid [1]). (*B*) A 70-year-old woman with false-negative CNN prediction. Transverse view of the left thyroid gland demonstrating a mixed cystic and solid nodule, CNN malignancy score of 0.014, and TIRADS 3 (three points, isoechoic [1], mixed cystic/solid [1], and macrocalcifications [1]). (*C*) A 47-year-old man with false-positive CNN prediction. Transverse view of the isthmus demonstrating a hypoechoic nodule containing calcifications, CNN malignancy score of 0.809, and TIRADS 4 (six points, very hypoechoic [3], solid [2], and macrocalcifications [1]). (*D*) A 61-year-old woman with false-positive CNN prediction. Sagittal view of the left thyroid gland demonstrating a hypoechoic nodule, CNN malignancy score of 0.886, and TIRADS 4 (five points, very hypoechoic [3] and solid [2]).

© 2020 Wolters Kluwer Health, Inc. All rights reserved.

Copyright © 2020 Wolters Kluwer Health, Inc. All rights reserved.

easily be drawn around the nodule at the same time with minimal additional effort. Further work could explore an automated thyroid nodule segmentation and measurement tool as a postprocessing step following initial image acquisition.

While an interesting and important future area of investigation, this work was not meant to be an exhaustive examination of multiple model architectures. A multitude of CNN architectures are available including many traditional as well as newer lightweight networks such as MobileNetV2,[40] ShuffleNetV2,[41] and NasNet,[42] which have demonstrated slightly better performance on very large data sets such as ImageNet. However, these improvements are modest at best (72.0% vs. 70.6% top-1 accuracy for MobileNetV1 and MobileNetV2, respectively[40]), and it is unclear if there would be any meaningful performance improvements with much smaller medical data sets. Further investigation of the optimal CNN architecture for thyroid nodule characterization could potentially lead to improvement classification performance.

In conclusion, this study demonstrates that a CNN can accurately classify and differentiate malignant from benign thyroid nodules on ultrasound imaging. The malignancy scores generated by our model could be used as additional clinical tool alongside current professional guideline such as TIRADS to risk stratify thyroid nodules when identifying those patients requiring further histopathologic workup.

REFERENCES

1. Dean DS, Gharib H. Epidemiology of thyroid nodules. *Best Pract Res Clin Endocrinol Metab*. 2008;22:901–911.
2. Nam-Goong IS, Kim HY, Gong G, et al. Ultrasonography-guided fine-needle aspiration of thyroid incidentaloma: correlation with pathological findings. *Clin Endocrinol (Oxf)*. 2004;60:21–28.
3. Smith-Bindman R, Lebda P, Feldstein VA, et al. Risk of thyroid Cancer based on thyroid ultrasound imaging characteristics: results of a population-based study. *JAMA Intern Med*. 2013;173:1788–1796.
4. Martinez-Tello FJ, Martinez-Cabruja R, Fernandez-Martin J, et al. Occult carcinoma of the thyroid. A systematic autopsy study from Spain of two series performed with two different methods. *Cancer*. 1993;71: 4022–4029.
5. Grant EG, Tessler FN, Hoang JK, et al. Thyroid ultrasound reporting lexicon: white paper of the ACR Thyroid Imaging, Reporting and Data System (TIRADS) committee. *J Am Coll Radiol*. 2015;12:1272–1279.
6. 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: 162–167.
7. Acharya UR, Swapna G, Sree SV, et al. A review on ultrasound-based thyroid cancer tissue characterization and automated classification. *Technol Cancer Res Treat*. 2014;13:289–301.
8. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. 2014. Available at: https://arxiv.org/abs/1409.0575. Accessed May 2, 2020.
9. Ren S, He K, Girshick R, et al. Faster r-cnn: towards real-time object detection with region proposal networks. In: *Advances in Neural Information Processing Systems*. 2015:91–99.
10. Jégou S, Drozdzal M, Vazquez D, et al. The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 I.E. Conference on IEEE. 2017;1175–1183. Available at: https://arxiv.org/abs/1611.09326. Accessed May 2, 2020.
11. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*. 2017;284:574–582.
12. Cicero M, Bilbily A, Colak E, et al. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. *Invest Radiol*. 2017;52:281–287.
13. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. 2017. Available at: https://arxiv.org/abs/1711.05225. Accessed May 2, 2020.
14. Prevedello LM, Erdal BS, Ryu JL, et al. Automated critical test findings identification and online notification system using artificial intelligence in imaging. *Radiology*. 2017;285:923–931.
15. Lee H, Tajmir S, Lee J, et al. Fully automated deep learning system for bone age assessment. *J Digit Imaging*. 2017;30:427–441.
16. Larson DB, Chen MC, Lungren MP, et al. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology*. 2018;287:313–322.
17. Kamnitsas K, Ledig C, Newcombe VF, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*. 2017;36:61–78.
18. Iandola FN, Han S, Moskewicz MW, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. 2016. Available at: https://arxiv.org/abs/1602.07360. Accessed May 2, 2020.
19. Howard AG, Zhu M, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. 2017. Available at: https://arxiv.org/abs/1704.04861. Accessed May 2, 2020.
20. Middleton WD, Teefey SA, Reading CC, et al. Multiinstitutional analysis of thyroid nodule risk stratification using the American College of Radiology Thyroid Imaging Reporting and Data System. *Am J Roentgenol*. 2017;208: 1331–1341.
21. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*. 2018;286:800–809.
22. Frates MC, Benson CB, Charboneau JW, et al. Management of thyroid nodules detected at US: Society of Radiologists in ultrasound consensus conference statement. *Radiology*. 2005;237:794–800.
23. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. 2015. Available at: https://arxiv.org/abs/1512.03385. Accessed May 2, 2020.
24. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowledge Data Eng*. 2010;22:1345–1359.
25. Cawley GC, Talbot NLC. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J Mach Learn Res*. 2010;11:2079–2107.
26. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM*. 2017;60: 84–90.
27. Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014. Available at: https://arxiv.org/abs/1412.6980. Accessed May 2, 2020.
28. LeDell E, Petersen M, van der Laan M. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electron J Stat*. 2015;9:1583–1607.
29. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7:91.
30. 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.
31. Gharib H, Papini E, Garber JR, 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: appendix. *Endocr Pract*. 2016;22:1–60.
32. 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;17:370–395.
33. 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;14:587–595.
34. Middleton WD, Teefey SA, Reading CC, et al. Comparison of performance characteristics of American College of Radiology TI-RADS, Korean Society of Thyroid Radiology TIRADS, and American Thyroid Association guidelines. *Am J Roentgenol*. 2018;210:1148–1154.

© 2020 Wolters Kluwer Health, Inc. All rights reserved.

Copyright © 2020 Wolters Kluwer Health, Inc. All rights reserved.

35. Ha EJ, Na DG, Baek JH, et al. US fine-needle aspiration biopsy for thyroid malignancy: diagnostic performance of seven society guidelines applied to 2000 thyroid nodules. *Radiology*. 2018;287:893–900.
36. Oda H, Miyauchi A, Ito Y, et al. Incidences of unfavorable events in the management of low-risk papillary microcarcinoma of the thyroid by active surveillance versus immediate surgery. *Thyroid*. 2016;26:150–155.
37. Liu YI, Kamaya A, Desser TS, et al. A Bayesian network for differentiating benign from malignant thyroid nodules using sonographic and demographic features. *Am J Roentgenol*. 2011;196:W598–W605.
38. Wu H, Deng Z, Zhang B, et al. Classifier model based on machine learning algorithms: application to differential diagnosis of suspicious thyroid nodules via sonography. *Am J Roentgenol*. 2016;207:859–864.
39. 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;30:477–486.
40. Sandler M, Howard A, Zhu M, et al. MobileNetV2: inverted residuals and linear bottlenecks. 2018. Available at: https://arxiv.org/abs/1801.04381. Accessed May 2, 2020.
41. Ma N, Zhang X, Zheng H-T, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design. 2018. Available at: https://arxiv.org/abs/1807.11164. Accessed May 2, 2020.
42. Zoph B, Vasudevan V, Shlens J, et al. Learning transferable architectures for scalable image recognition. 2017. Available at: https://arxiv.org/abs/1707.07012. Accessed May 2, 2020.

© 2020 Wolters Kluwer Health, Inc. All rights reserved.

Copyright © 2020 Wolters Kluwer Health, Inc. All rights reserved.