ISSN: 2577-2996

PROGRESS IN MEDICAL SCIENCES, 2024
VOL 8, NO. 2, PAGE 1 – 5
DOI: doi.org/10.47363/PMS/2024(8)203

RESEARCH ARTICLE

Open Access

Radiologists Versus Artificial Intelligence in Distinguishing Between Thyroid Nodules on Ultrasound Images

Karima Bahmane[1]*, AKSASSE HAMID[1], Brahim Alkhalil Chaouki[1] and Soukaina Wakrim[2]

[1]Systems engineering and decision support laboratory, National School of Applied Sciences Agadir, Morocco

[2]Department of Radiology, University Hospital Center, Agadir, Morocco

ABSTRACT

Introduction: In order to distinguish between benign and malignant thyroid nodules on ultrasound pictures, we created three convolutional neural network (CNN) models as well as ensemble models. We then evaluated the diagnostic efficacy of CNN models against that of two radiologists.

Material and Methods: Between 2020 and 2022, we analyzed ultrasound pictures of 100 individuals who had 120 thyroid nodules that were verified by surgical pathology. In a test set, two radiologists used ultrasound scans to retroactively diagnose benign and malignant thyroid nodules. Using 80 and 40 thyroid nodule ultrasound images, respectively, three CNNs (ResNet50, DenseNet12, and VGGNet) were tested and trained-validated. Next, for the model ensemble, we choose the two models that performed the best diagnostically on the test set. Then, a comparison was made between the integrated model and the diagnostic performance of two radiologists utilizing three CNN models.

Results: 50 of the 120 thyroid nodules were benign, and 70 were malignant. For the diagnosis of thyroid cancer, two radiologists under the curves (AUCs) ranged from 0.659 to 0.754. The three CNN models and the ensemble model had AUCs ranging from 0.801 to 0.907 for the diagnosis of thyroid cancer. AUC differences were statistically substantial ($p < 0.05$) between the CNN models and the radiologists' models. With the highest AUC score was the ensemble model.

Conclusions: When it came to using ultrasonography to differentiate between benign and malignant thyroid nodules, three CNN models and an ensemble model outperformed radiologists. The ensemble model's diagnostic performance demonstrated good promise and continued to improve.

ARTICLE HISTORY

Received April 17, 2024
Accepted April 22, 2024
Published April 30, 2024

KEYWORDS

Thyroid, Ultrasound, Artificial Intelligence, Image Classification, CNN

Introduction

There are restrictions with conventional ultrasonography. Ultrasound is more sensitive to image artifacts and the patient's position than other imaging modalities. Certain illnesses, like adenoma and nodular goiter, cannot be diagnosed using ultrasound due to its low specificity.

There is a high degree of subjectivity in the diagnostic outcomes, which makes the variations between physicians at different diagnostic levels inevitable. Furthermore, the workload for doctors who perform thyroid ultrasound scanning is rising due to the rise in patients and the time-consuming nature of the scan.

Artificial intelligence has been applied to the diagnosis of medical imaging as computer science has advanced. The most popular method, image classification, creates an intelligent classification model based on the various elements in ultrasound images and produces a precise diagnosis.

Typically, this technology involves multiple phases, such as pre-processing images, extracting features, and classifying data.

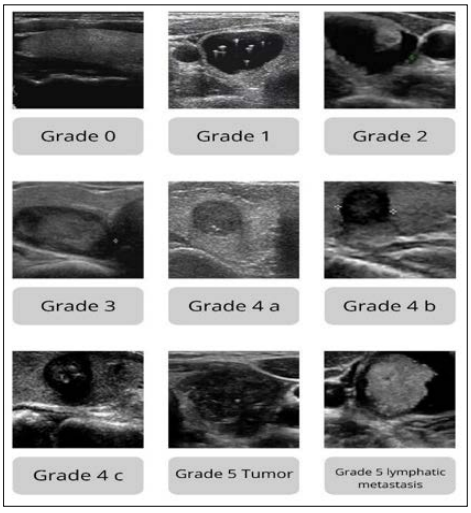

Figure 1: The classification standard for TI-RADS system. ACR TI-RADS Grade 4, where the benign and malignant nodules can easily overlap, in this modified TI-RADS, was further divided into grade 4a, 4b, and 4c.

Contact: Karima Bahmane, Systems Engineering and Decision Support Laboratory, National School of Applied Sciences Agadir, Morocco.

Purpose of the Study

An AI-assisted approach could greatly enhance radiologists diagnostic abilities and assist in lowering the quantity of needless fine needle aspirations for thyroid nodules. Our results suggest that AI diagnostic programs be implemented in thyroid nodule management clinical settings.

Table 1: Malignant Risk of Thyroid Nodules Reported by Modified TI-RADS and Corresponding Medical Suggestion

Modified TI-RADS classification	Definition	Risk of malignancy	Recommended
TI-RADS 2	benign lesions	0	Long-term follow-up
TI-RADS 3	high probability of benignity	<5%	Short-term follow-up
TI-RADS 4a	possible benignity	5~15%	FNA
TI-RADS 4b	high probability of malignancy	15~90%	FNA
TI-RADS 5	highly suggestive of malignancy	>90%	Clinical treatment

Material and Methods

Patient Data

Between 2020 and 2022, 200 patients with thyroid nodules from Agadir's Hassan II Hospital were retrospectively enrolled.

The following were the inclusion criteria: A standard ultrasound examination was performed prior to the biopsy, there was no prior surgical treatment or FNA biopsy, and the ultrasound image quality satisfies diagnostic criteria and calibration analysis. Uncertain diagnostic findings on histology served as the exclusion criterion.

Next, these nodules were divided into test and training-validation sets at random.

Image acquisition and preprocessing 80 Images from a training-verification set comprised the entire 120 cases that were randomly divided. There were 40 cases in the test set. Random rotation and random horizontal flipping were applied as picture augmentation techniques for the training set's ROI. All ROI images were scaled to 224 × 224 × 3 and normalized to pixels between 0 and 1. They were also modified to a window width of 350 and a window level of 40.

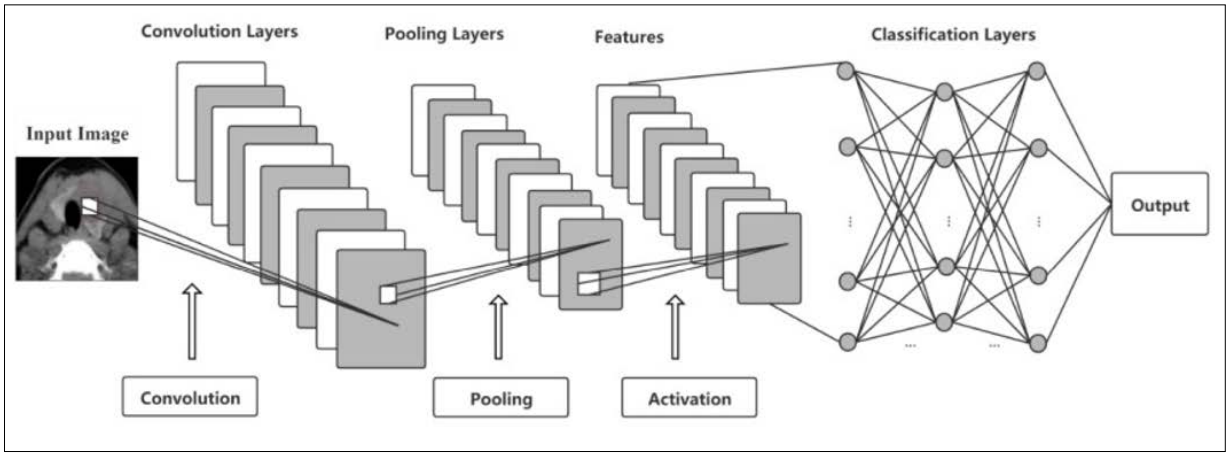

Figure 2: Basic architecture of convolutional neural network (CNN) for image classification problems

Model Training-Validation and Testing, Ensemble Model

CNN's fundamental design is depicted in Figure 2. Based on preoperative ultrasound pictures, three deep learning CNN models were chosen to distinguish between benign and malignant thyroid nodules. The CNN On the training validation set, fivefold crossvalidation was carried out by each model. In models ResNet50, DenseNet121, and VGGNet were employed. On ImageNet, all networks embraced the pre-trained models. ImageNet is an image database arranged using the WordNet structure, with hundreds or thousands of photographs representing each node in the network. In terms of scale and diversity, ImageNet is greater. training, 30 iterations was the maximum. With a batch size of four and an optimizer named Adam, the learning rate started at 5e-5 and decreased to the 9th power of the number of iterations.

The model with the greatest AUC on the validation set was chosen and tested on the test set for each model's 5-fold cross-validation.

We combined the anticipated outcomes of each model's folds on the test set, identified two models with superior diagnostic performance, and ultimately produced the ensemble model of the two models.

Performance Evaluation

The area under the receiver operating characteristic curve (ROC), sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV) of the test dataset were used to evaluate the performances of the three CNN models as well as the ensemble model. CNN models and radiologists are compared. Based only on the ultrasound images of the test set, two radiologists classified each thyroid nodule as benign or cancerous, while remaining blind to the FNA histology results. Their diagnostic abilities were contrasted with those of the ensemble model and the three CNN models. Lesion detection and heat map of attention to get insight into CNN's interpretation of ultrasound images for thyroid nodule classification, we retrieved the final convolution layer from the training model prior to the fully connected layer's classification.

Results

Patient Characteristics

50 (41.6%) benign nodules and 70 (58.3%) malignant nodules were found. These nodules were divided into a test set (15 benign and 20 malignant nodules) and a training-validation set (35 benign and 50 malignant nodules) at random. Between benign and malignant thyroid nodules, there was no discernible difference in the mean size of the nodules, the female-to-male ratio, or the age of the patients ($p > 0.05$).

Table 2: The Diagnostic Performances of 3 Convolutional Neural Network (CNN) Models and 2 Radiologists on the Test Set

	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
ResNet50	0.945	0.874	0.837	0.911	0.877	0.872
DenseNet121	0.943	0.869	0.884	0.866	0.833	0.898
VGGNet	0.901	0.808	0.872	0.768	0.740	0.878
Radiologist J	0.587	0.586	0.593	0.580	0.520	0.644
Radiologist S	0.754	0.748	0.802	0.705	0.677	0.705

PPV: Positive Predictive Value

NPV: Negative Predictive Value

Radiologist Junior: Inexperienced Radiologists

Radiologist Senior: Experienced Radiologists

Comparisons between the 2 Radiologists and CNN Models for Malignant and Benign Thyroid Nodules

There was good agreement as evidenced by the intra- and inter-class correlation coefficients (ICCs), which were 0.961 and 0.768, respectively.

Table 3 displays the two radiologists' diagnoses for the ultrasound images from the test set.

It should come as no surprise that Radiologist Senior, an experienced radiologist, performed noticeably better than Radiologist Junior, a novice radiologist. Table 3 displays the comparison findings of the three models as well as the ensemble model with Radiologist J and Radiologist S. When it came to diagnosing benign and malignant thyroid nodules, the three models and the combined model outperformed Radiologist J by a significant margin ($p > 0.05$). There were substantial differences ($p < 0.05$) in AUC between the three models, the ensemble model, and Radiologist S.

Resnet50, Densenet121, VGG Net, and Radiologist S differed significantly in terms of specificity (0.911, 0.866, 0.884, 0.857 vs. 0.705, respectively; $p = 0.000$, 0.005, 0.001, and 0.009, respectively).

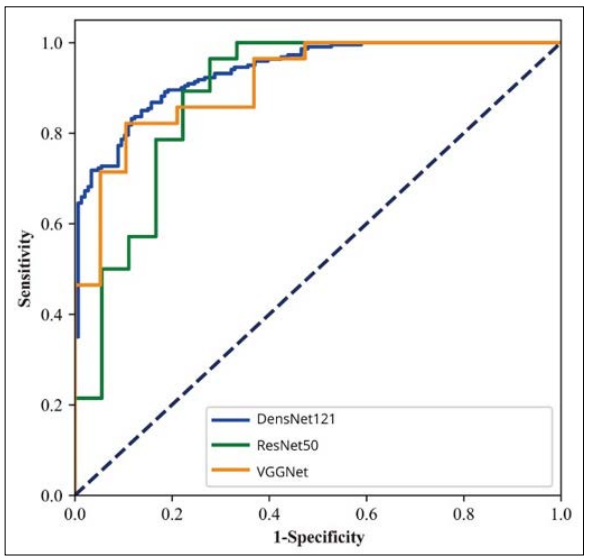

Figure 3: Receiver operating characteristic curves. The receiver operating characteristic curves of the 3 convolutional neural network (CNN) models on the test set

Resnet50, Densenet121, VGGNet, the ensemble model, and Radiologist S all differed significantly in accuracy (0.874, 0.869, 0.859, 0.859, 0.859 vs. 0.7475, respectively; p = 0.001, 0.002, 0.005, 0.005, and 0.005, respectively). Radiologist S and the ensemble model differed significantly in PPV (0.920 vs. 0.705; p = 0.042). Resnet50, Densenet121, VGGNet, and Radiologist S all differed significantly from one another in terms of NPV (0.877, 0.833, 0.845, 0.822 vs. 0.677, respectively; p = 0.002, 0.012, 0.008, and 0.021).

Ultimately, the ensemble model and the three CNN models outperformed the radiologists.

Table 3: Comparisons of diagnostic performances between 5 convolutional neural network (CNN) models and an ensemble model for malignant and benign thyroid nodules

	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
RN50 vs. Radiologist J	< 0.001*	< 0.001*	0.001*	< 0.001*	< 0.001*	< 0.001*
DN121 vs. Radiologist J	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*
VGGNet vs. Radiologist J	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*
IM vs. Radiologist J	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*
RN50 vs. Radiologist S	< 0.001*	0.001*	0.839	< 0.001*	0.002*	0.321
DN121 vs. Radiologist S	< 0.001*	0.002*	0.263	0.005*	0.012*	0.119
VGGNet vs. Radiologist S	< 0.001*	0.005*	0.824	0.001*	0.008*	0.361
IM vs. Radiologist S	< 0.001*	0.005*	0.064	0.053	0.056	0.042*
Radiologist J vs. Radiologist S	< 0.001*	0.001*	< 0.001*	0.001*	0.624	0.006*

Radiologist Junior: Inexperienced Radiologists

Radiologist Senior: Experienced radiologists

*Represent Statistically Significant (p < 0.05)

Discussion

This research has some advantages over the previous studies. Firstly, whereas the research mentioned above only employed one model, we trained three models in all. Second, we performed pairwise comparisons across all models in addition to analyzing each model's diagnostic performance for benign and malignant thyroid nodules.

While all three models produced good results, ResNet50 and DenseNet121 performed better in terms of diagnosis.

Lastly, two models with superior diagnostic capabilities for the ensemble model were also chosen in this investigation.

There are various restrictions on this study. First, papillary thyroid cancer accounted for about 83.2% of the malignant thyroid nodules in this study, which could explain why malignant thyroid nodules appear too uniformly on ultrasound.

To increase the number of different types of thyroid nodules that are malignant, follow-up examinations are required.

Second, in order to verify the research's diagnostic performance and generalizability, an external validation study and an increased sample size were necessary. This study was based on a single hospital and had a modest total sample size of 120 ultrasound images.

Lastly, the areas of interest sketch used in this investigation was created manually by a different radiologist, which limits its clinical applicability and is not sufficiently automatic. It will be further enhanced to do an automatic or semi-automatic drawing in the following phase [1-18].

Conclusion

In summary, radiologists were outperformed by three models and the ensemble model in differentiating between benign and malignant thyroid nodules on ultrasound images.

The ensemble model's diagnostic performance outperformed the single model and demonstrated good promise. Consequently, CNN can be used as a helpful technique to differentiate between benign and malignant thyroid nodules and to classify between the 5 stages of TI-RADS.

References

[1] Russ G, Leboulleux S, Leenhardt L, Laszlo H. Thyroid incidentalomas: epidemiology, risk stratification with ultrasound and workup. Eur Thyroid J. 2014; 3: 154-163.

[2] Angell TE, Maurer R, Wang Z, M I Kim, C A Alexander, et al. A Cohort Analysis of Clinical and Ultrasound Variables Predicting Cancer Risk in 20,001 Consecutive Thyroid Nodules. J Clin Endocrinol Metab. 2019; 104: 5665-5672.

[3] Hoang JK, Branstetter BF, Gafton AR, Wai KL, Christine MG. Imaging of thyroid carcinoma with CT and MRI: approaches to common scenarios. Cancer Imaging. 2013; 13: 128-139.

[4] Shie P, Cardarelli R, Sprawls K, Kimberly G F, Alan T. Systematic review: prevalence of malignant incidental thyroid nodules identified on fluorine-18 fluorodeoxyglucose positron emission tomography. Nucl Med Commun. 2009; 30: 742-748.

[5] Brito JP, Gionfriddo MR, Al Nofal A, Kasey RB, Aaron LL, et al. The accuracy of thyroid nodule ultrasound to predict thyroid cancer: systematic review and meta-analysis. J Clin Endocrinol Metab. 2014; 99: 1253-1263.

[6] Haugen BR, Alexander EK, Bible KC, Gerard MD, Susan JM, et al. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid. 2015; 26: 1-133.

[7] Gharib H, Papini E, Garber JR, Daniel SD, R MH, et al. AACE/ACE/AME Task Force on Thyroid Nodules. American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi medical guidelines for clinical practice for the diagnosis and management of thyroid nodules-2016 update. Endocr Pract. 2016; 22: 622-639.

[8] Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015; 61: 85-117.

[9] Shen D, Wu G, Suk HI. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng. 2017; 19: 221-248

[10] Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015; 33: 831-838.

[11] Aerts HJ, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014; 5: 4006.

[12] Ko SuY, Lee JiH, Yoon JH, Na H , Hong E, et al. Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound. Head Neck. 2019; 41: 885-891,

[13] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comp Vis. 2015; 115: 211-252.

[14] Bolei Z, Aditya K, Agata L, Aude O, Antonio T. Learning Deep Features for Discriminative Localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Seattle, USA, 2016 June 27-30.

[15] Zhu Y, Fu Z, Fei J. An image augmentation method using convolutional network for thyroid nodule classification by transfer learning. In Proceedings of the 3rd IEEE International Conference on Computer and Communication (1819-1823) Chengdu, China, 13-16 December 2017.

[16] Chi J, Walia E, Babyn P, Wang J, Groot G, et al. Thyroid Nodule Classification in Ultrasound Images by Fine-Tuning Deep Convolutional Neural Network. J Digit Imaging. 2017; 30: 477-486.

[17] Song W, Li S, Liu Ji, Qin H, Zhang B, et al. Multitask Cascade Convolution Neural Networks for Automatic Thyroid Nodule Detection and Recognition. IEEE J Biomed Health Inform. 2019; 23: 1215-1224.

[18] Nguyen DT, Kang JK, Pham TD, Batchuluun G, Park RK, et al. Ultrasound Image Based Diagnosis of Malignant Thyroid Nodules Using Artificial Intelligence. Sensors (Basel). 2020; 20: 1822.