Automated machine learning-based radiomics analysis versus deep learning-based classification for thyroid nodule on ultrasound images: a multi-center study

Zelong Liu
Icahn School of Medicine at Mount Sinai
New York, NY, USA
zelong.liu@icahn.mssm.edu

Louisa Deyer
Dwight School
New York, NY, USA
2025ldeyer@dwight.edu

Arnold Yang
Westview High School
Portland, OR, USA
arnold.tianyi.yang@gmail.com

Steven Liu
New York University
New York, NY, USA
swl8539@nyu.edu

Jingqi Gong
Icahn School of Medicine at Mount Sinai
New York, NY, USA
jingqi.gong.beta@gmail.com

Yang Yang
Icahn School of Medicine at Mount Sinai
New York, NY, USA
yang.yang@mssm.edu

Mingqian Huang
Mount Sinai Hospital
New York, NY, USA
Mingqian.huang@mountsinai.org

Amish Doshi
Mount Sinai Hospital
New York, NY, USA
amish.doshi@mountsinai.org

Meng Lu
University of Massachusetts Boston
Boston, MA, USA
Ivan.lu.usa@gmail.com

Denise Lee
Mount Sinai Hospital
New York, NY, 10029, USA
Denise.Lee@mountsinai.org

Timothy Deyer
East River Medical Imaging
Cornell Medicine
New York, NY, USA
tdeyer@eastriverimaging.com

Xueyan Mei
Icahn School of Medicine at Mount Sinai
New York, NY, USA
xueyan.mei@icahn.mssm.edu

***Abstract*—Often, the characteristics of thyroid nodules need to be determined by fine needle aspiration (FNA) biopsy. The increasing applications of machine learning and deep learning algorithms provide alternative noninvasive methods to study thyroid nodules on ultrasound images. Many studies examined the feasibility of convolutional neural networks or radiomics feature extraction to analyze the characteristics of thyroid nodules. In this study, we built an automated radiomics analysis system by combining thyroid segmentation via U-Net and radiomics feature extraction. Our proposed machine learning-based automated radiomics analysis was compared to a deep learning-based convolutional neural network method in a two-center thyroid nodule classification task. It is shown that the automated radiomics analysis can accurately segment thyroid nodules to facilitate clinical diagnosis by achieving dice scores of 0.77 and 0.74 on internal and external sets respectively. In addition, the proposed automated radiomics analysis approach can improve sensitivity, negative predictive value (NPV) and positive predictive value (PPV) by 352.1%, 3.9% and 187.5% respectively, while reducing the false negative rate by 46.6%.**

***Index Terms*—radiomics, thyroid nodule, deep learning, ultrasound, radimagenet**

I. Introduction

Thyroid nodules are commonly detected on thyroid ultrasound. Studies have shown that 10-15% of thyroid nodules are malignant [1]. However, the ultrasound characteristics of thyroid nodules can be sometimes difficult to classify so a fine needle aspiration (FNA) biopsy is often performed to study the histopathology of the nodules to direct management. The Bethesda grading system [2] is a pathology guideline for classifying thyroid nodule FNA histopathology. The diagnostic system grades range from 2-6 and represent an increasing risk of malignancy. Grade 2 is equivalent to a benign lesion (0-3% probability of malignancy) while Grade 6 indicates a malignant lesion (97-99% risk of malignancy). Grades 3 and 4 are intermediate and often rely on additional molecular testing to determine management. The FNA is an invasive procedure with attendant risk of infection and bleeding. Non-invasive approaches such as radiomic analysis and deep learning-based classification could be used to diagnose benign and malignant thyroid nodules on ultrasound images thereby decreasing the need for an invasive procedure.

Radiomics [3] features are high-level textural features ex-

2471-7819/22/$31.00 ©2022 IEEE
DOI 10.1109/BIBE55377.2022.00013

tracted from medical images that might not be detected by visual inspection. Extracting the high level features requires both the original medical image as well as an image with an annotated region of interest (ROI). Radiomics features include the analysis of first-order, second-order, and high-order statistical features to study the shape, size, and border. Prior studies have used radiomics to predict BRAF mutation in papillary thyroid carcinoma [4, 5], to predict lymph node metastasis in patients with thyroid cancer [6], to predict thyroid malignancy [7, 8] on thyroid ultrasound images. Typical radiomics features are extracted from expert's annotations. To make this an automated system, we proposed to use the U-Net [9] method to automatically segment the nodules. U-Net is a convolutional network designed for fast and accurate segmentation on medical images. It consists of downsampling and upsampling processes to learn what and where the important information is located.

In addition to radiomics analysis, deep learning is another popular approach that has been implemented to classify thyroid nodules. Convolutional neural networks (CNN) based deep learning methods have shown promising results for detecting malignant thyroid nodules [10]. Both radiomics analysis and CNN-based methods are non-invasive but require different levels of annotation. Radiomics analysis needs pixel-level annotation of the ROI for further textual feature extraction, whereas CNN-based method only needs image-level labels. To our knowledge, comparing automated radiomics analysis to CNN for classification of thyroid nodules in multi-center studies has not been done before.

In this study, we compare an automated radiomics analysis to CNN-based thyroid nodule classification in a multi-center study. The automated radiomics system includes two steps: 1) a U-Net network to segment nodules; 2) radiomics feature extraction based on the predicted masks. Comparison between these two approaches were analyzed in AUC, accuracy, sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), and false negative rate (FNR). Fig.1 illustrates the design of this study.

II. Related Works

The ultrasound images are critical to identify thyroid nodules, but can require fine needle aspiration to accurately categorize a nodule as benign or malignant. In recent years, machine learning and deep learning algorithms have greatly improved the classification of thyroid nodules in order to reduce the need for fine needle aspiration. To directly analyze ultrasound images, both convolutional neural networks and radiomics were used in many studies. In one study, Mei et al. [11] extracted image features via autoencoders, local binary patterns, and histogram of oriented gradient descriptors and built machine learning models to predict the benignity of thyroid nodules. Zhu et al. [12] found that deep neural networks can achieve better performance than conventional machine learning algorithms. Another important application of deep learning in thyroid nodule analysis is the use of U-Net to identify the region of the nodule in the ultrasound image. The success of this method of nodule detection was shown in a study by Ying et al [13].

There are also studies applying radiomics in medical imaging. Park et al. [8] applied radiomics to improve the risk stratification system of thyroid malignancy. Radiomics and convolutional neural networks were compared by Truhn et al. [14] in the classification of breast cancer malignancy. They found CNN outperformed radiomics.

These previous works show that CNN models and radiomics for the classification of thyroid nodules have been widely implemented. However, studies that compared both approaches on multi-center studies have not been performed.

III. Materials and Method

A. Study cohort

We collected 3,176 images from 626 nodules between April 2017 and December 2019 from an outpatient radiology facility in New York (Center A). All patients had a diagnostic ultrasound and FNA with cytopathologic results, which were used as the ground truth to train deep learning and machine learning models. The patients had a median age of 58 (IQR: 46 - 68) and 475 (75.9%) were females. 2,435 of the 3,176 images were benign (Bethesda 2), 527 were Bethesda grade 3 and 4, 264 were Bethesda grade 5 and 6. In this study, we considered bethesda grade 3-6 as non-benign nodules. Of the 626 nodules from Center A, 80% (501 nodules with 2,585 images; 1,976 benign images; 400 benign nodules) were randomly split into the training set and 20% (125 nodules with 591 images; 455 benign images; 100 nodules) were used as the validation set.

A total of 190 nodules and 950 images including 173 benign nodules from an external dataset (Center B) [15] were used for the test set. The test cohort data were collected between April 2017 and May 2018. In the test cohort, the median age was 57 (IQR: 45 - 68) and 157 (82.6%) were females. 173 of the nodules were benign.

Table I reports the patient's demographics and distribution of labels among the training, validation, and test sets. Each nodule in our internal dataset contained 5 images on average while the external dataset consisted of thyroid ultrasound cine-clips. To have a similar distribution of data collected from Center A, we selected 5 images from the cine-clip that contained the largest 5 ROIs.

TABLE I
Patient demographics.

	Train (n=501)	Validation (n=125)	Test (n=190)
Age	[a]58 (46-68)	[a]58 (46-66)	[a]57 (45-68)
Female	382 (76.2%)	93 (74.4%)	157 (82.6%)
Male	119 (23.8%)	32 (25.6%)	33 (17.4%)
Benign	400 (79.8%)	100 (80.0%)	173 (91.1%)
Non-benign	101 20.2%)	25 (20.0%)	17 (8.9%)

[a]Data in parenthesis indicates IQR.

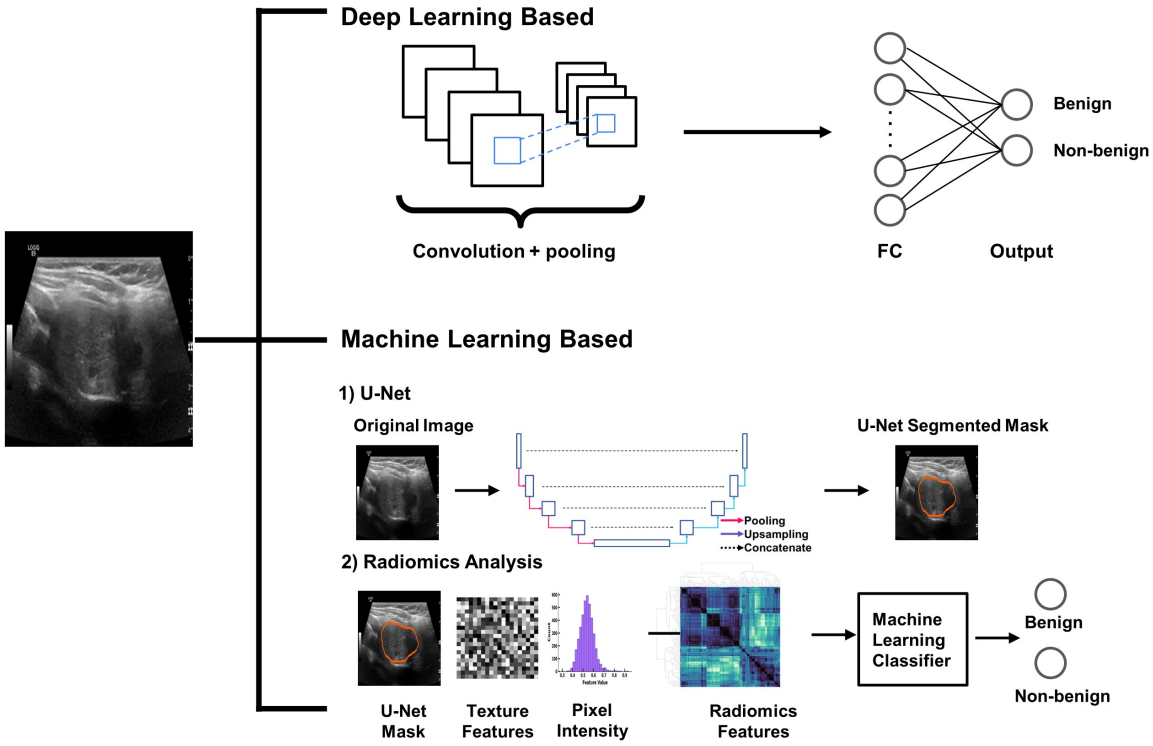

Fig. 1. Overview of this study.

B. U-Net

Radiomics analysis requires the inputs of a delineated ROI that needs an expert's annotation. The dataset from Center 1 was labeled by a senior musculoskeletal and interventional radiologist with 12-years of clinical experience. To make this an automated process, a U-Net was developed to automatically segment the thyroid nodules on ultrasound images. The predicted masks generated by the U-Net were used for radiomics analysis and feature extraction.

The architecture of U-Net contained a downsampling path and a upsampling path united with skip connection. In the downsampling process, the convolutional layers extracted and encoded image features into deep features. Then, the upsampling path decoded the deep features and each decoder level was fused with the corresponding high resolution encoder level to produce a pixel level semantic segmentation mask.

We used the DenseNet121 [16] architecture as the encoding pathway of U-Net. A total of five layers of DenseNet 121 were used: input_1, conv1/relu, pool2_relu, pool3_relu, pool4_relu and relu. Pretrained weights from the RadImageNet [17] DenseNet121 model were imported to initialize the encoders. Learning rates of 0.001 and 0.0001, optimizers of Adam and RMSprop, and batch size of 16 and 32 were tested to find the best U-Net model.

Dice score was used as the metric to report the performance of U-Net as compared to the expert's ground truth. The U-Net model that achieved the highest dice score on the validation set was appointed to generate masks for radiomics analysis.

C. Radiomics feature extraction and machine learning classifiers

Radiomics is a quantitative analysis to study characteristics of medical images. We used the pyradiomics [18] package in python to extract textural features from predicted masks generated by the U-Net model. A total of 932 features were extracted. Excluding descriptive features, 168 features were first order (i.e., mean of pixels); 220 were gray level co-occurrence matrix (GLCM) features; 160 were gray level run length matrix (GLRLM) features; 160 were gray level size zone matrix (GLSZM) features; 140 were gray-level dependence matrix (GLDM) features; and 50 were neighboring gray tone difference matrix (NGTDM) features. Pearson correlation analysis was performed on the radiomics features. Features showing a correlation score greater than 0.8 were dropped. This resulted in 883 features to feed into machine learning classifiers. We used XGboost [19], random forest [20], and multi-layer perceptron classifiers [21] for the binary classification of thyroid nodules. The classifiers were evaluated in metrics of area under the receiver operating characteristic curve (ROC AUC), sensitivity, specificity, NPV, PPV, and FNR, where the calculations of these metrics were displayed in (1) - (5), where TP, FN, TN, and FP stands for true positives, false negatives, true negatives, and false positives.

$$Sensitiviy = \frac{TP}{TP + FN} \tag{1}$$

$$Specificity = \frac{TN}{TN + FP} \quad (2)$$

$$NPV = \frac{TN}{TN + FN} \quad (3)$$

$$PPV = \frac{TP}{TP + FP} \quad (4)$$

$$FNR = \frac{FN}{TP + FN} \quad (5)$$

D. Deep learning-based classification

We evaluated DenseNet121, InceptionV3 [22], and EfficientNetB4 [23] backbones for the classification of thyroid nodules. Pretrained weights developed on the RadImageNet database were imported for transfer learning. The original ultrasound images were resized to 512 x 256 (width x height) pixels. Adam optimizer with a learning rate of 0.001, and a batch size of 16 were used to train the CNN models. All models were trained for 40 epochs and the best model demonstrating the lowest validation loss was used for evaluation on the external test set and comparisons to machine learning-based approaches. The gradient class activation map (Grad-CAM) [24] was used to visualize features learned by the best CNN model. Key features were highlighted in red colors. The Grad-CAM was also compared to the U-Net model regarding model interpretability of radiomics-based model and deep learning-based model.

IV. Results

A. Thyroid nodule segmentation via U-Net for radiomics feature extraction

The automated radiomics analysis of thyroid nodules consists of two steps. The first step is predicting the segmentation mask of the thyroid nodule using U-Net, and the second step is extracting radiomics features using the predicted mask. Our U-Net model utilized the pretrained weights from RadImageNet which is a large public radiology database with pre-trained CNN weights for medical imaging research. The RadImageNet pretrained weights were implemented for the downsampling encoder of the U-Net to extract deep features of thyroid nodules. During the fine tuning process of U-Net, the model with optimizer RMSprop, learning rate 0.001 and batch size 16 achieved the highest dice score on the internal validation set (0.77), and best dice score (0.70) based on the test set.

The predicted masks of all thyroid images, including both the internal dataset and external dataset, were generated by this best U-Net model. These masks were then used for the image feature extraction via Pyradiomics. We built three machine learning models, Random Forest, XGBoost, and MLP to predict histopathologic outcomes based on 883 radiomic features. In Table II, Random Forest, XGBoost, and MLP achieved AUC scores of 0.59, 0.60, and 0.65 on the internal validation set, and 0.55, 0.54, and 0.61 on the external test set, respectively. MLP achieved the highest AUC on the validation set, and was thereby appointed for the comparison of CNN models.

TABLE II
Comparison of AUC between CNN and radiomics.

	Internal dataset	External dataset
CNN DenseNet121	0.57	0.38
CNN InceptionV3	0.54	0.57
CNN EfficientNetB4	0.61	0.39
Radiomics Random Forest	0.59	0.55
Radiomics XGBoost	0.60	0.54
Radiomics MLP	**0.65**	**0.61**

B. CNN for thyroid non-benignity prediction

Compared to the automated radiomics analysis, we applied three convolutional neural network architectures, DenseNet121, InceptionV3 and EfficientNetB4, and conducted transfer learning with their RadImageNet pretrained weights. In Table II, DenseNet121, InceptionV3 and EfficientNetB4 achieved AUC scores of 0.57, 0.54 and 0.61 on the internal validation set, and 0.38, 0.57 and 0.39 on the external test set, respectively. EfficientNetB4, given its best performance on the validation set, was chosen for comparisons with the radiomics approach.

C. Comparison between the best CNN model and the best automated radiomics model

We used the Youden index to find the optimal threshold for the ROC curve analysis of the best CNN model and the best automated radiomics model. The receiver operating characteristics curves (ROC) comparing the EfficientNetB4 model and the MLP model were plotted in Fig. 2. The MLP demonstrated higher AUC on both validation and test sets. This showed that the MLP model that trained on radiomics features extracted from ROIs had better generalizability than the EfficientNetB4 model that used the whole image as an input which included extraneous image data. In addition, the Youden index of EfficientNetB4 was 0.079 from the validation set and the sensitivity, specificity, NPV, PPV and FNR was 11.7%, 72.2%, 89.3%, 4.0%, 88.2%, respectively (Table III). For the best automated radiomics model, the Youden index of MLP was 0.108 from the validation set and the sensitivity, specificity, NPV, PPV and FNR was 52.9%, 60.1%, 92.8%, 11.5% and 47.1%, respectively. Except for specificity, the radiomics model illustrated better results in sensitivity, NPV, PPV, and FNR.

Moreover, the U-Net model that learned from expert's annotations showed a more consistent interpretability than the CNN model. Fig. 3 showed an example of a benign nodule that were both correctly predicted by the radiomics model and the CNN model. The colorbar indicated the probability of being non-benign in red. However, the U-Net model showing the predicted segmented nodule (orange) was close enough to the ground truth (blue), whereas the Grad-CAM suggested that CNN picked random pixels outside the ground truth. The

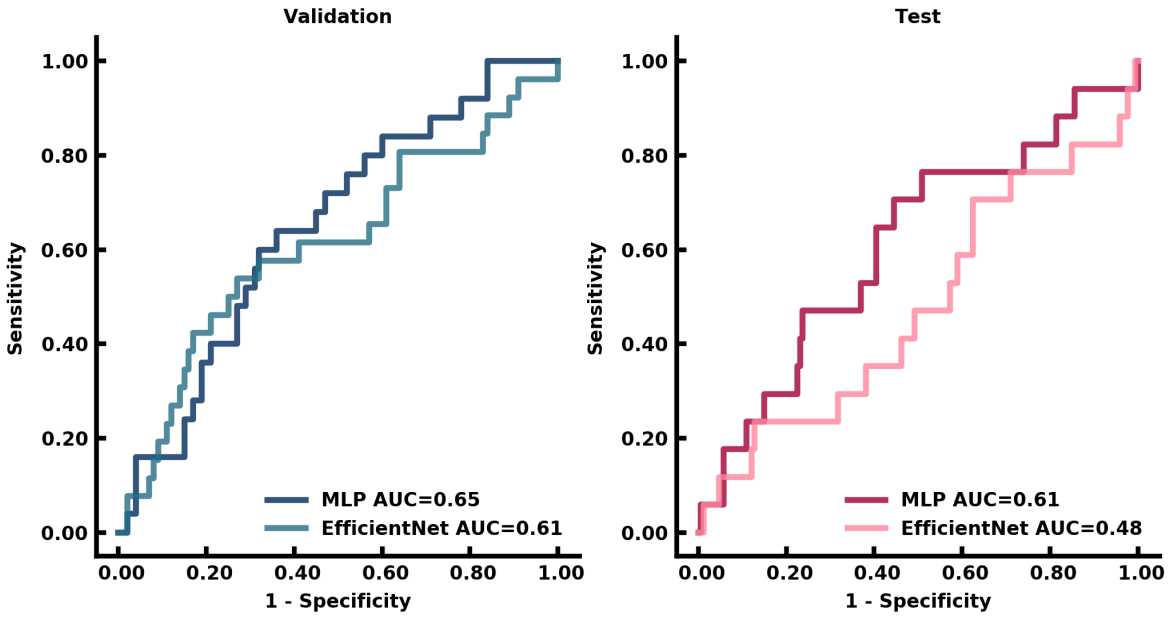

Fig. 2. ROC curves of CNN and radiomics models

TABLE III
COMPARISON BETWEEN BEST CNN AND AUTOMATED RADIOMICS MODELS.

	EfficientNetB4 (CNN)	U-Net + MLP (radiomics)
Sensitivity	11.7%	**52.9%**
Specificity	72.2%	60.1%
NPV	89.3%	**92.8%**
PPV	4.0%	**11.5%**
FNR	88.2%	**47.1%**
AUC	0.48	**0.61**

decoding pathway of the U-Net provided more accurate information of where the ROI was, which could further improve the radiomics analysis to extract accurate textual features.

V. CONCLUSION

Machine learning-based radiomics analysis is better than deep learning-based classification of thyroid nodules on ultrasound images. Radiomics analysis that requires the input of a region of interest can significantly improve sensitivity, NPV and PPV by 352.1%, 3.9% and 187.5% respectively, and reduce the false negative rate by 46.6%. Our proposed automated radiomics approach combining U-Net and MLP can improve the efficiency of annotating the nodules and thus result in more accurate diagnosis.

ACKNOWLEDGMENT

Xueyan Mei, PhD and Timothy Deyer, MD are co-corresponding authors of this manuscript.

REFERENCES

[1] S. C. Kamran, E. Marqusee, M. I. Kim, M. C. Frates, J. Ritner, H. Peters, C. B. Benson, P. M. Doubilet, E. S. Cibas, J. Barletta, N. Cho, A. Gawande, D. Ruan, F. D. Moore, K. Pou, P. R. Larsen, and E. K. Alexander, "Thyroid nodule size and prediction of cancer," The Journal of Clinical Endocrinology & Metabolism, vol. 98, no. 2, pp. 564–570, 2013.

[2] M. A. Melo-Uribe, Á. Sanabria, A. Romero-Rojas, G. Pérez, E. J. Vargas, V. Gutiérrez, and M. C. Abaúnza, "The Bethesda System for reporting thyroid cytopathology in Colombia: Correlation with histopathological diagnoses in oncology and non-oncology institutions," Journal of Cytology, vol. 32, no. 1, p. 12, 2015.

[3] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images are more than pictures, they are data," Radiology, vol. 278, no. 2, pp. 563–577, 2016.

[4] Y. Cao, X. Zhong, W. Diao, J. Mu, Y. Cheng, and Z. Jia, "Radiomics in differentiated thyroid cancer and nodules: Explorations, application, and limitations," Cancers, vol. 13, no. 10, p. 2436, 2021.

[5] M.-r. Kwon, J. H. Shin, H. Park, H. Cho, S. Y. Hahn, and K. W. Park, "Radiomics study of thyroid ultrasound for predicting BRAF mutation in papillary thyroid carcinoma: Preliminary results," American Journal of Neuroradiology, vol. 41, no. 4, pp. 700–705, 2020.

[6] F. Li, D. Pan, Y. He, Y. Wu, J. Peng, J. Li, Y. Wang, H. Yang, and J. Chen, "Using ultrasound features and radiomics analysis to predict lymph node metastasis in patients with thyroid cancer," BMC Surgery, vol. 20, no. 1, 2020.

[7] M. Gul, K.-J. C. Bonjoc, D. Gorlin, C. W. Wong, A. Salem, V. La, A. Filippov, A. Chaudhry, M. H. Imam, and A. A. Chaudhry, "Diagnostic utility of Radiomics in thyroid and head and neck cancers," Frontiers in Oncology, vol. 11, 2021.

[8] V. Y. Park, E. Lee, H. S. Lee, H. J. Kim, J. Yoon, J. Son, K. Song, H. J. Moon, J. H. Yoon, G. R. Kim, and J. Y. Kwak, "Combining radiomics with ultrasound-based risk stratification systems for thyroid nodules: An approach for improving performance," European Radiology, vol. 31, no. 4, pp. 2405–2413, 2020.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Lecture Notes in Computer Science, pp. 234–241, 2015.

[10] X. Li, S. Zhang, Q. Zhang, X. Wei, Y. Pan, J. Zhao, X. Xin, C. Qin, X. Wang, J. Li, F. Yang, Y. Zhao, M. Yang, Q. Wang, Z. Zheng, X. Zheng,

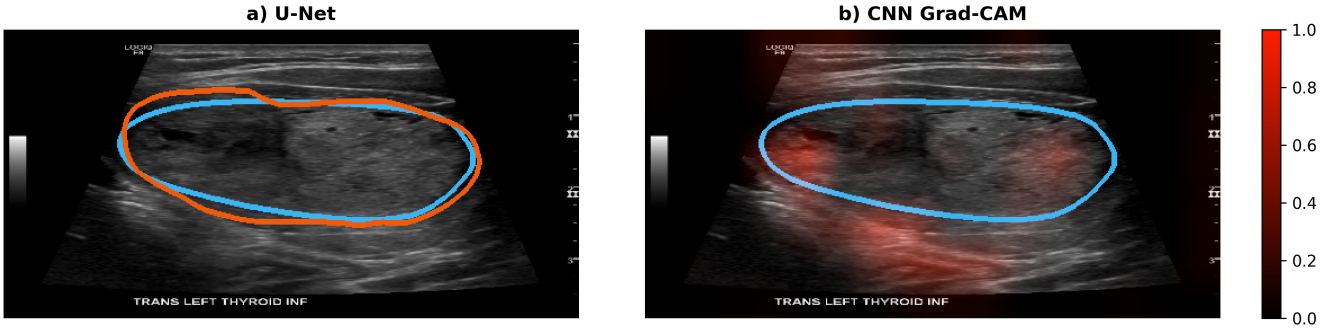

Fig. 3. Comparisons between the U-Net and CNN Grad-CAM of a benign thyroid nodule. a) The segmented nodule by U-Net (orange outline indicates the predicted ROI while blue was ground truth). b) Grad-CAM of the EfficientNetB4 model, where highlighted orange regions were the important pixels as selected by the CNN model in making predictions.

X. Yang, C. T. Whitlow, M. N. Gurcan, L. Zhang, X. Wang, B. C. Pasche, M. Gao, W. Zhang, and K. Chen, "Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: A retrospective, multicohort, Diagnostic Study," The Lancet Oncology, vol. 20, no. 2, pp. 193–201, 2019.

[11] X. Mei, X. Dong, T. Deyer, J. Zeng, T. Trafalis and Y. Fang, "Thyroid Nodule Benignty Prediction by Deep Feature Extraction," 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 241-245, 2017.

[12] Y. Zhu, Q. Sang, S. Jia, Y. Wang, and T. Deyer, "Deep Neural Networks could differentiate Bethesda class III versus class IV/V/VI," Annals of Translational Medicine, vol. 7, no. 11, pp. 231–231, 2019.

[13] X. Ying, Z. Yu, R. Yu, X. Li, M. Yu, M. Zhao, and K. Liu, "Thyroid nodule segmentation in ultrasound images based on cascaded convolutional neural network," Neural Information Processing, pp. 373–384, 2018.

[14] D. Truhn, S. Schrading, C. Haarburger, H. Schneider, D. Merhof, and C. Kuhl, "Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast MRI," Radiology, vol. 290, no. 2, pp. 290–297, 2019.

[15] R. Yamashita, T. Kapoor, M. N. Alam, A. Galimzianova, S. A. Syed, M. Ugur Akdogan, E. Alkim, A. L. Wentland, N. Madhuripan, D. Goff, V. Barbee, N. D. Sheybani, H. Sagreiya, D. L. Rubin, and T. S. Desser, "Toward reduction in false-positive thyroid nodule biopsies with a deep learning–based risk stratification system using US CINE-clip images," Radiology: Artificial Intelligence, vol. 4, no. 3, 2022.

[16] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261-2269, 2017.

[17] X. Mei, Z. Liu, P. M. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, C. Cao, K. E. Link, T. Yang, Y. Wang, H. Greenspan, T. Deyer, Z. A. Fayad, and Y. Yang, "RadImageNet: An open radiologic deep learning research dataset for effective transfer learning," Radiology: Artificial Intelligence, 2022.

[18] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, "Computational radiomics system to decode the radiographic phenotype," Cancer Research, vol. 77, no. 21, 2017.

[19] T. Chen and C. Guestrin, "XGBoost," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[20] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[21] P. M. Atkinson and A. R. Tatnall, "Introduction neural networks in remote sensing," International Journal of Remote Sensing, vol. 18, no. 4, pp. 699–709, 1997.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818-2826, 2016.

[23] M. Tan, and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2019 International conference on machine learning, PMLR, pp. 6105-6114, 2019.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618-626, 2017.