Deep Learning Analysis of Thyroid Nodule Ultrasound Images has High Sensitivity and Negative Predictive Value to Rule-out Thyroid Cancer

Nikita Pozdeyev. *University of Colorado*

Toshimasa Clark[1], Trevor Cohen[2], Bryan R. Haugen[1], Devika Subramanian[3], Nikita Pozdeyev[1], Manjiri Dighe[2], Martin Barrio[1], Michael G. Leu[2]
[1]University of Colorado Anschutz Medical Campus, Aurora, Colo., [2]University of Washington, Seattle, Wash., [3]Rice University, Houston, Texas

Purpose: To evaluate deep learning analysis of thyroid nodule ultrasound images as a rule-out test for thyroid malignancy.

Methods: Supervised deep learning (DL) classifier of thyroid nodules was trained on 32,545 thyroid US images from 621 nodules representing all major benign and malignant types of thyroid lesions and tested on an independent set of 145 nodules collected at a different healthcare system in the United States. The Big Transfer BiT-M ResNet-50x1 convolutional neural net architecture was modified to contain 3, 4, 6 and 3 PreActBottleneck units per block 1 through 4. Weights pretrained on the ImageNet-21k dataset were loaded and weights for blocks 3 and 4 were fine-tuned for the binary classification task of distinguishing benign and malignant thyroid nodules.

Results: The deep learning thyroid nodule classifier achieved an area under receiver operating characteristic curve (AUROC) of 0.889 on five-fold cross-validation. The AUROC improved when images were scaled by nodule size and six randomly selected cine clip frames were added to the training set per epoch. GradCAM class activation heatmaps revealed that microcalcifications and spongiform appearance were reliably recognized by the classifier as malignant and benign features, respectively. Spongiform nodules were found to be benign even when microcystic spaces constituted less than 50% of nodule volume. To investigate the clinical relevance of the benign vs. malignant classifier, the binary classification threshold for the probability of malignancy generated by model was set at 7% to achieve sensitivity and negative predictive value (NPV) comparable to that of the fine needle aspiration biopsy (FNA). At this threshold, cross-validated deep-learning model achieved a sensitivity of 90%, specificity of 63%, positive predictive value (PPV) of 46% and negative predictive value of 94%. When tested on an independent image set that includes 18 classic papillary thyroid cancers (PTC), 5 follicular variant PTC, 4 medullary thyroid cancers, 3 follicular thyroid cancers (FTC), and 1 Hurthle cell thyroid cancer, the DL classifier achieved AUROC of 0.88, sensitivity of 97%, specificity of 61%, PPV of 40% and NPV of 99%. A single minimally-invasive FTC that had no suspicious features on thyroid ultrasound was incorrectly classified as benign.

Conclusions: This study demonstrates that the ultrasound-based deep-learning classifier of thyroid nodules achieves sensitivity and negative predictive value comparable to that of thyroid fine needle aspiration (FNA). Clinicians may use this tool to augment clinical judgment when determining whether to perform FNA procedures.

Presentation Type: Poster
Presentation Date: Saturday, June 11
Presentation Time: 1-3 PM
Location: ENDOExpo
Rapid Fire Poster Presentation: Saturday, June 11 from 1-2 PM