A Multi-view Deep Learning-based Architecture for Thyroid Nodules Detection and Characterization

Sanaz Vahdati, MD, Mayo Clinic Rochester; Bardia Khosravi, MD, MPH, MHPE; Pouria Rouzrokh, MD, MPH, MPHPE; Mana Moassefi, MD; Zeynettin Akkus, PhD; Bradley J. Erickson, MD, PhD, CIIP, FSIIM

Introduction
Thyroid cancer is regarded as one of the most rapidly growing malignancies in the general population, with an incidence rate of 14.6 per 100,000 in the United States. Thyroid Ultrasound (US) is the primary method to evaluate thyroid nodules and determine whether a nodule requires Fine Needle Aspiration (FNA), an invasive and costly procedure.

Hypothesis
We propose a Deep Learning (DL) based pipeline to detect and classify thyroid nodules into benign or malignant groups relying on two views of US imaging.

Methods
Transverse and longitudinal views from B-mode US images of single thyroid nodules from 1003 patients were collected retrospectively. A total of 101 cases were held out as a testing set, and the rest of the data was used in five-fold cross-validation (CV). Two You Look Only Once (YOLO) v5 models, one for each view, were trained to detect nodules and classify them as benign or malignant. For each view, five models were developed during the CV, one for each fold, which was ensembled by using non-max suppression (NMS) to boost their collective performance. Finally, an extreme gradient boosting machine (XGBoost) model was trained on outputs of the ensembled models for both transverse and longitudinal views to yield a final prediction of malignancy status for a single nodule. Mean average precision at 50% overlap (mAP0.5) was used to evaluate the YOLO models. The whole pipeline’s performance was evaluated using Area under receiving operating curve (AUROC). All the metrics were reported on the holdout test set.

Results
The total number of malignant nodules was 518/1003 (51%), with 50 malignant nodules in the test set. The averaged mAP0.5 of the five cross-validation folds for transverse, and longitudinal YOLO models were 0.70 (SD: 0.038) and 0.72 (SD: 0.039), respectively. The ensembled models for each view achieved a mAP0.5 of 0.797 (for the transverse view) and 0.716 (for the longitudinal view). The whole pipeline, created by the XGBoost ensemble, reached an AUROC of 0.84 (CI 95%: 0.75-0.91) with sensitivity and specificity of 0.84 and 0.627, respectively. The negative predictive value of the pipeline was 0.799, and the F1 score of 0.756.

Conclusion
We demonstrate the robustness of an ensemble of DL models receiving multiple views of thyroid US images to detect and characterize thyroid nodules.

Statement of Impact
Our DL model can assist radiologists in better evaluation of thyroid nodules.
Figure 1. Diagram of the multi-view deep learning architecture with two ensembles, YOLOV5-NMS and XGBoost, for detection and characterization of the thyroid nodule.

Figure 2. Five-Fold cross validation for Transverse and Longitudinal YOLOV5 model

Figure 3. Receiving Operating Curve (ROC) of ensembled model. AUC = Area Under ROC Curve.

**Keywords**
Deep learning; Thyroid nodules; Ultrasound; Artificial Intelligence; YOLOV5