Interim Report

Deep Learning of Thyroid Nodule Detection and Classification

Name: Zhou Zihan, 3035947843

Supervisor: Dr. Kenneth K.Y. Wong

Data: 19/01/2024

1. Introduction

Thyroid cancer is the most frequent endocrine malignancy, with an incidence rate of approximately 14.6 per 100,000 individuals in the United States, predominantly affecting women, where it ranks as the seventh most prevalent form of cancer [1]. Consequently, there is a growing requirement for trustworthy imaging methods to examine thyroid nodules, with ultrasound (US) imaging being the generally used modality.

Due to the widespread use of ultrasonography, many incidental thyroid nodules are found during unrelated examinations. However, only about 5% of thyroid nodules are proven to be malignant [2]. In 2017, the ACR promulgated the TI-RADS system based on ultrasound, which standardized the nodular classification and reporting of ultrasound examinations to ensure that each nodule was classified and a definitive diagnostic suggestion was given (biopsy, follow-up, or no intervention), improving specificity [2]. Many studies have found that using the TI-RADS system could reduce the false negative rate of malignant nodules. However, the consistency of judgment between different radiologists using the TI-RADS system to classify nodules still does not meet clinical requirements, posing a significant challenge [3]. The strong subjectivity of manual ultrasound examination and the large number of unnecessary biopsies highlight the need for improved diagnostic tools for nodular examination [3].

Deep learning models can mitigate the subjective factors involved in diagnosis. Machine learning has provided valuable insights and decision support in analyzing medical images in radiology settings [4]. Some studies have demonstrated that the diagnostic ability of deep learning can match or surpass that of doctors [5]. However, most studies on thyroid ultrasound using deep learning do not adequately address the need for both effective localization and classification of nodules. This project aims to conduct a systematic model selection process to identify the most effective algorithms for the object detection or instance segmentation task of thyroid nodules in ultrasound images, as well as for the classification tasks. Building on existing literature and techniques in deep learning, the project will propose

a robust deep-learning-based pipeline that is able to localize the position of thyroid nodules from US images, and then classify the malignant risk of nodules based on the feature extracted from the region of interest (ROI).

2. Methodology

The proposed methodology introduces an automated thyroid nodule analysis system (Figure 1) with two main components: nodule detection and TI-RADS classification. The detection phase utilized YOLO11n for automatic nodule extraction, while the classification phase employed multiple CNN models. The methodology was implemented using the Stanford AIMI Dataset, with data preprocessing and feature extraction optimized for medical imaging requirements.

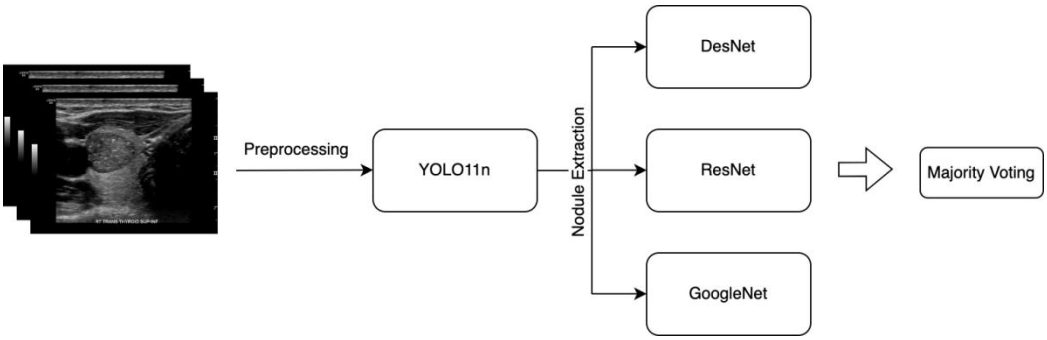

Figure 1. Thyroid nodule detection and classification pipeline architecture overview.

2.1. Data Collection and Sources

This study primarily utilizes the Stanford AIMI Thyroid Ultrasound Cine-clip Dataset [6], which comprises 167 patients with biopsy-confirmed thyroid nodules from Stanford University Medical Center. The dataset contains 192 nodules captured in approximately 18,000 ultrasound frames. Additionally, through collaboration with Queen Mary Hospital, approximately 100 de-identified scans with expert annotations will be acquired. The Stanford AIMI shared dataset serves as the primary training source due to its comprehensive nature and sufficient size for deep learning model training. The Queen Mary Hospital dataset will be subsequently used for model fine-tuning to fit real-world clinical use.

2.2. Data Analysis and Characteristics

Characteristic	Stanford AIMI Dataset	
	Benign	**Malignant**
Age (y)	56.8 ± 15.2	48.3 ± 14.1
Sex - Female	144 (82.3%)	15 (88.2%)
Sex - Male	31 (17.7%)	2 (11.8%)
TI-RADS Level		
1	1 (0.6%)	0 (0.0%)
2	10 (5.7%)	0 (0.0%)
3	52 (29.7%)	0 (0.0%)
4	78 (44.6%)	5 (29.4%)
5	34 (19.4%)	12 (70.6%)
Total	175	17

Table 1. Clinical characteristics of the Stanford AIMI Dataset.

Table 1 presents the clinical characteristics of the Stanford AIMI Dataset, categorizing patients based on benign and malignant thyroid nodules. The dataset analysis reveals several key characteristics. The mean age for benign cases (56.8 ± 15.2 years) is notably higher than malignant cases (48.3 ± 14.1 years). Gender distribution shows a predominance of female patients in both categories, with 82.3% (144) and 88.2% (15) in benign and malignant groups respectively. The TI-RADS classification demonstrates a distinct pattern, with benign nodules distributed across all levels (1-5), predominantly in levels 3 (29.7%) and 4 (44.6%). In contrast, malignant nodules are concentrated in higher TI-RADS levels, with 70.6% at level 5 and 29.4% at level 4. These characteristics highlight significant class imbalance in the dataset, with 175 benign cases substantially outnumbering 17 malignant cases. This imbalance will need to be carefully addressed in the model development phase to ensure effective classification performance for both categories.

2.3. Image Preprocessing Strategies

Data augmentation is a widely used technique in deep learning that enhances the diversity of the training dataset through various image transformations [7]. While common augmentation methods include rotations, distortions, noise addition, and brightness/contrast modifications,

medical image processing requires careful consideration to maintain diagnostic integrity. In this study, only horizontal flipping (mirroring) was implemented on the Stanford AIMI Shared Dataset. This conservative approach was chosen to preserve the anatomical and pathological characteristics of thyroid nodules while effectively expanding the training data. More aggressive augmentation techniques were avoided to prevent potential distortions that might compromise diagnostic accuracy.

2.4. Thyroid Nodule Detection Model Development

The You Only Look Once (YOLO) model was selected for thyroid nodule detection due to its proven track record in medical imaging applications [8]. YOLO models are particularly well-suited for medical image interpretation, offering superior accuracy and minimal background errors compared to traditional detection approaches. The specific implementation uses YOLO11n, the latest iteration in the YOLO series, which has been optimized for precise object localization through bounding box prediction.

During training, the model processes Stanford AIMI Thyroid Ultrasound Dataset images through its detection architecture. Each image is paired with annotations in normalized coordinates (center_x, center_y, width, height) following YOLO's format. The model was trained with systematically evaluated hyperparameters, with key parameters including learning rate, batch size, and learning rate scheduler. The dataset is first split with 15% reserved as the test set. The remaining data undergoes 5-fold cross-validation, systematically dividing it into five equal portions. This ensures each subset serves alternately as validation data while the remaining portions form the training set, maximizing data utilization while enabling robust evaluation of model generalization.

2.5. Thyroid Nodule Cancer Risk Classification Model System

2.5.1. Image Cropping Methods

Thyroid ultrasound ROIs extracted from ground truth labels were used as input for subsequent networks. However, due to varying nodule sizes in the Stanford AIMI dataset and the fixed input size requirement (224 × 224 pixels) of classification networks, a standardized

image cropping method is essential for preprocessing thyroid ultrasound images shown in Figure 2. The implemented cropping method centers each Region of Interest (ROI) within a fixed-size square patch, applying zero-padding where necessary. This process maintains the original scale and proportions of nodules while preserving the critical TI-RADS features: composition, echogenicity, shape, margin, and punctate echogenic foci. Each ultrasound image undergoes the same standardization process to ensure consistent dimensions for network input. By preserving the nodule's original proportions and key diagnostic features, the method enables accurate TI-RADS assessment while ensuring compatibility with deep learning architectures.

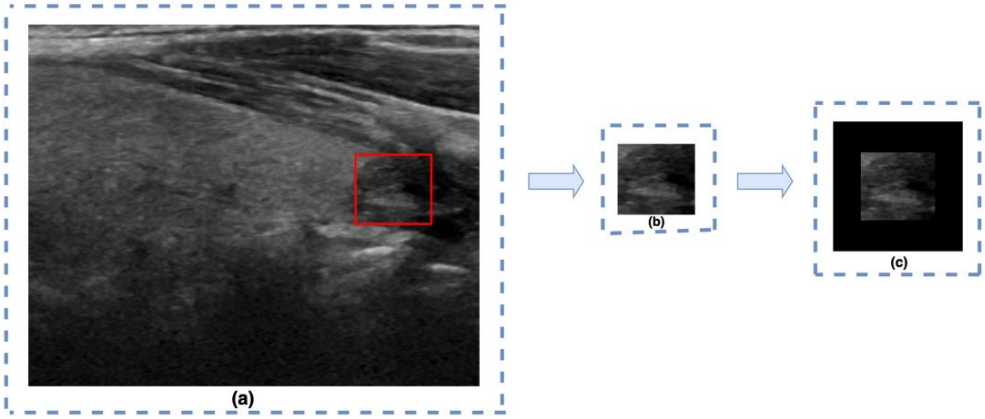

Figure 2. Image cropping pipeline: (a) Target nodule marked by ground truth label; (b) Region of Interest (ROI) extracted from ultrasound image; (c) image cropping method

2.5.2. Classification Model Development

Supervised deep learning algorithms are trained using labeled data, where the correct label or class for each image is known in advance. This approach is particularly effective for analyzing thyroid ultrasound images, as it delivers accurate and consistent detection and classification of nodules [9]. Several convolutional neural network architectures will be evaluated, including VGG16, VGG19, and ResNet50, leveraging their deep feature learning capabilities. Those classification models directly learn thyroid nodule appearance patterns from ROI images to predict ACR-TIRADS risk levels, focusing on critical characteristics: composition, echogenicity, shape, margin, and punctate echogenic foci. This deep learning

approach enables automatic feature learning from the image data, eliminating the need for manual feature extraction and allowing the model to capture complex visual patterns associated with different TI-RADS categories.

The training process will utilize labeled images from the standardized Stanford AIMI thyroid ultrasound database, employing an 80-20 split for training and validation in a 5-fold cross-validation scheme. Regularization techniques and network visualization methods will be used to analyze learned features and their contribution to classification decisions. Model performance will be rigorously evaluated across multiple validation sets using metrics such as accuracy and F1-score. The implemented strategy is expected to yield a model capable of accurate and reliable cancer risk classification for clinical screening and diagnosis support.

3. Results and Discussion

This section presents interim results from the evaluation of the YOLO-based thyroid nodule detection system and preliminary design of the classification model through comprehensive experiments. The performance assessment includes evaluation metrics and cross-validation results to demonstrate the detection model's capability in automatic ROI extraction from ultrasound images. Additionally, technical challenges encountered during model optimization and data preprocessing are discussed, along with their proposed solutions. This section ends with a discussion of the remaining work plan for the project.

3.1. Thyroid Nodule Detection Model Analysis

3.1.1. Evaluation Metrics

To assess the effectiveness of the YOLO model, several standard object detection metrics were used, including precision, recall, mean Average Precision at IoU threshold 0.5 (mAP50), and mean Average Precision across IoU thresholds from 0.5 to 0.95 (mAP50-95). These metrics provide complementary insights: precision and recall measure detection accuracy and completeness, while mAP values evaluate localization quality at different overlap thresholds. A 5-fold cross-validation strategy was employed to ensure robust evaluation.

3.1.2. Performance Evaluation

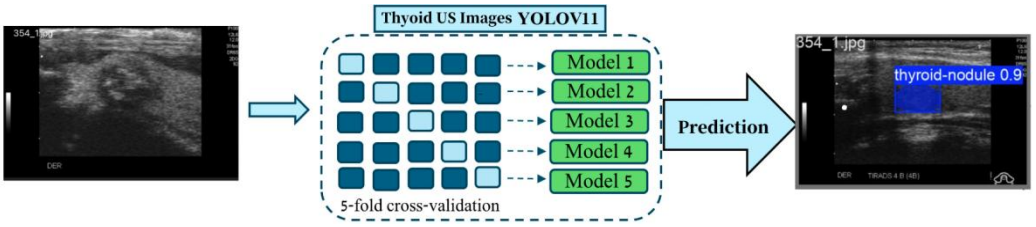

Figure 3. Achitecture of YOLO11n for thyroid nodule detection.

The performance of the YOLO-based detection model was evaluated on the reserved test set from Stanford AIMI Shared Dataset, as demonstrated in Figure 3. The model could identify the presence of thyroid nodules within ultrasound images and generate bounding boxes for nodule localization, with each bounded region representing a detected nodule.

Fold	Precision	Recall	mAP50	mAP50-95
1	0.875	0.560	0.735	0.501
2	0.929	0.424	0.668	0.473
3	0.895	0.552	0.739	0.472
4	0.902	0.529	0.729	0.489
5	0.877	0.609	0.760	0.472
Mean ± SD	0.896 ± 0.021	0.535 ± 0.068	0.726 ± 0.034	0.481 ± 0.013

Table 2. Cross-validation results of YOLO detection performance across five folds.

The detection performance was systematically evaluated across all five folds during the testing process, as shown in Table 2. The results demonstrated consistent detection capabilities across folds, with average metrics of precision = 0.896 ± 0.021, recall = 0.535 ± 0.068, mAP50 = 0.726 ± 0.034, and mAP50-95 = 0.481 ± 0.013. Notable variations in performance were observed, with fold 5 achieving the highest mAP50 at 0.760, and fold 2 reaching the highest precision at 0.929.

Despite the challenges in detecting small or ambiguous nodules in ultrasound images, the

model demonstrated reliable performance in most cases. The high precision scores (>0.87 across all folds) indicate the reliability in the detected nodules, suggesting minimal false positive detections. The relatively lower recall values suggest room for improvement in detecting all present nodules, this could potentially be addressed by adjusting the confidence threshold based on specific clinical requirements. The stable mAP50 scores above 0.72 (except fold 2) demonstrate robust performance in accurate localization, while the mAP50-95 values indicate moderate performance across stricter IoU thresholds. These results suggest that the YOLO-based detection model could achieve satisfactory performance in automatic ROI extraction, potentially balancing detection accuracy with practical clinical requirements.

3.2. Thyroid Nodule Cancer Risk Classification Model Analysis

3.2.1. Preliminary Design and Results

The model is still in the development stage, and is expected to take the ground truth nodule regions as input and output their corresponding TI-RADS levels, providing an automated risk assessment for each identified nodule. Based on preliminary testing with VGG16, VGG19, and ResNet50 architectures on a small subset of the data, the classification model is expected to achieve accuracy rates between 65-75% for TI-RADS level prediction.

Several challenges have been identified during the initial classification model development. The significant class imbalance in the dataset, particularly for TI-RADS levels 1 and 5, may lead to a model bias towards the majority classes and poor recognition of the minority classes. The limited number of malignant cases (only 17 cases) could further exacerbate this issue. To mitigate these challenges, customized loss functions such as weighted cross-entropy or focal loss can be implemented to assign higher penalties for misclassification of minority classes, thereby encouraging the model to pay more attention to underrepresented samples.

3.3. Development Plan and Schedule

The project remains on schedule, with the detection model successfully implemented, demonstrating strong potential for identifying thyroid nodules in ultrasound images. This forms a solid foundation for the next phase of the project. The upcoming work will focus on

the training and refinement of the cancer risk classification model. Once completed, the classification model will be integrated with the detection stage to build a fully functional, end-to-end pipeline. This pipeline will enable both the detection of thyroid nodules and the assessment of their cancer risk levels, providing a streamlined solution for clinical use. The final pipeline is expected to be a robust and efficient tool, capable of reducing manual workloads and delivering objective, consistent insights to assist radiologists in diagnosis. The integration of detection and classification into a single system will ensure seamless operation, paving the way for future real-world deployment in clinical settings.

Reference List

[1] S. Vahdati, Z. Mortezapour, M. R. Arjmandi Taba, S. A. Motlagh, A. M. Nasr-Esfahani, E. Aboutalebi, H. Soltanian-Zadeh, and A. Mansour, "A Multi-View deep learning model for thyroid nodules detection and characterization in ultrasound imaging," Bioengineering, vol. 6, no. 2, p. 48, Jun. 2024. doi: 10.3390/bioengineering11070648

[2] D. Fresilli, L. Profili, and E. Venanzi, "Thyroid nodule Characterization: How to assess the malignancy risk. Update of the literature," Diagnostics, vol. 11, no. 8, p. 1374, Jul. 2021. doi: 10.3390/diagnostics11081374

[3] G. Low, Y. Ge, S. Finger, and C. Truong, "Tips for improving consistency of thyroid nodule interpretation with ACR TI-RADS," Journal of Ultrasonography, vol. 22, no. 88, pp. 51–56, Feb. 2022. doi: 10.15557/jou.2022.0009

[4] T.-C. Chang, "The role of computer-aided detection and diagnosis system in the differential diagnosis of thyroid lesions in ultrasonography," Journal of Medical Ultrasound, vol. 23, no. 4, pp. 177–184, Dec. 2015. doi: 10.1016/j.jmu.2015.10.002

[5] Y. Liu, Y. Feng, L. Qian, Z. Wang, and X. Hu, "Deep learning diagnostic performance and visual insights in differentiating benign and malignant thyroid nodules on ultrasound images," Experimental Biology and Medicine, vol. 248, no. 24, pp. 2538–2546, Dec. 2023. doi: 10.1177/15353702231220664

[6] "Stanford AIMI shared datasets," Oct. 15, 2024. https://stanfordaimi.azurewebsites.net/datasets/a72f2b02-7b53-4c5d-963c-d7253220bfdd5, accessed Oct. 15, 2024.

[7] H. Hassan, S. J. Choi, S. Periyasamy, S. K. Das, M. K. Ngwe, and Y. H. Kim, "Review and classification of AI-enabled COVID-19 CT imaging models based on computer vision

tasks," Computers in Biology and Medicine, vol. 141, p. 105123, Dec. 2021. doi: 10.1016/j.compbiomed.2021.105123

[8] D. Das, M. S. Iyengar, M. S. Majdi, J. J. Rodriguez, and M. Alsayed, "Deep learning for thyroid nodule examination: a technical review," Artificial Intelligence Review, vol. 57, no. 3, Feb. 2024. doi: 10.1007/s10462-023-10635-9

[9] S. Jung, H. Heo, S. Park, S.-U. Jung, and K. Lee, "Benchmarking deep learning models for instance segmentation," Applied Sciences, vol. 12, no. 17, p. 8856, Sep. 2022. doi: 10.3390/app12178856