Final Report

Deep Learning for Thyroid Nodule Detection and Classification

Name: Zhou Zihan, 3035947843

Supervisor: Dr. Kenneth K.Y. Wong

Data: 20/04/2025

Acknowledgment

I extend my sincere gratitude to Dr. Kenneth Wong and his Research Assistants for their guidance throughout this project.

Abstract

Accurate detection and classification of thyroid nodules in ultrasound images remain challenging tasks in clinical practice, with growing demand for automated assistance to reduce radiologist workload and improve diagnostic consistency. This study aimed to develop an automated system for thyroid nodule analysis, focusing on establishing a detection pipeline and implementing TI-RADS classification. A YOLO-based detection model was implemented using the Stanford AIMI Dataset (192 nodules), incorporating image preprocessing techniques and rigorous five-fold cross-validation for model evaluation. The detection model achieved a precision of 0.896 ± 0.021 and mAP50 of 0.726 ± 0.034, demonstrating reliable nodule localization across the dataset. A ResNet-101 multi-task classifier, applied to the ROIs cropped by YOLO, outputs discrete predictions and confidence scores for each of the five TI-RADS features (composition, echogenicity, shape, margin and foci); after majority-voting aggregation, the system flags TI-RADS 5 nodules (highest-risk) with a sensitivity of 0.835 ± 0.04 and specificity of 0.658 ± 0.03. These two models are integrated into a lightweight, responsive web application featuring drag-and-drop multi-frame uploads, real-time YOLO + ResNet inference and interactive TI-RADS visualization for seamless clinical integration.

Table of Content

List of Figures

List of Tables

Abbreviations

ACR - American College of Radiology

CNN - Convolutional Neural Network

IoU - Intersection over Union

mAP - mean Average Precision

mAP50 - mean Average Precision at 0.5 IoU threshold

mAP50-95 - mean Average Precision across IoU thresholds from 0.5 to 0.95

ROI - Region of Interest

SD - Standard Deviation

TI-RADS - Thyroid Imaging Reporting and Data System

US - Ultrasound

VGG - Visual Geometry Group (Neural Network)

YOLO - You Only Look

ResNet - Residual Network

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]. Nodule classification models based on CNN and ResNet50 have been proven to be effective and developed [5]. Also, the model classification ability of VGGNet has been found to be reliable as well [6]. Scholars have proposed that ensemble learning is a stable method for image classification, believing that ensemble models are superior to single models in malignant detection [7]. Again, some studies have demonstrated that the diagnostic ability of deep learning can match or surpass that of doctors [8]. 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).

An end-to-end automated thyroid nodule analysis system is developed. This report provides a comprehensive methodology of the data collection and annotation, as well as the model training methodology for detecting thyroid nodules and classifying malignancy risk groups in ultrasound images (Section 2). Lastly, the report discusses performance evaluation metrics, results, and addresses difficulties encountered (Section 3). This report also discusses implementation challenges, system limitations and lays out the roadmap for prospective clinical validation and deployment (Section 4).

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 YOLO11 for automatic nodule extraction, while the classification phase employed ResNet101 model. The methodology was implemented using the Stanford AIMI Dataset, with data preprocessing and feature extraction optimized for medical imaging requirements.

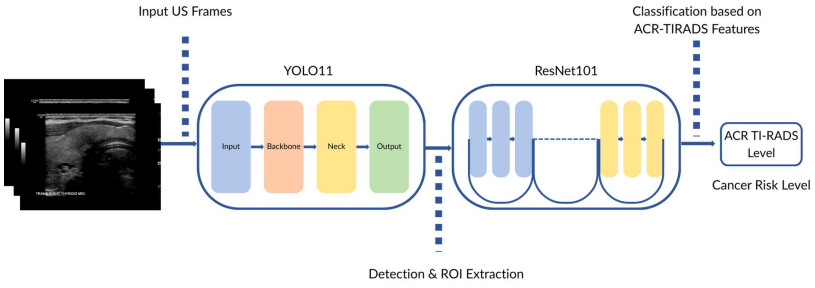

Figure 1. Thyroid nodule detection and classification pipeline architecture overview.

2.1. Data Collection and Sources

This study primarily utilizes the Stanford AIMI Thyroid Ultrasound Cine-clip Dataset [9], 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 [10]. 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 [11]. 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

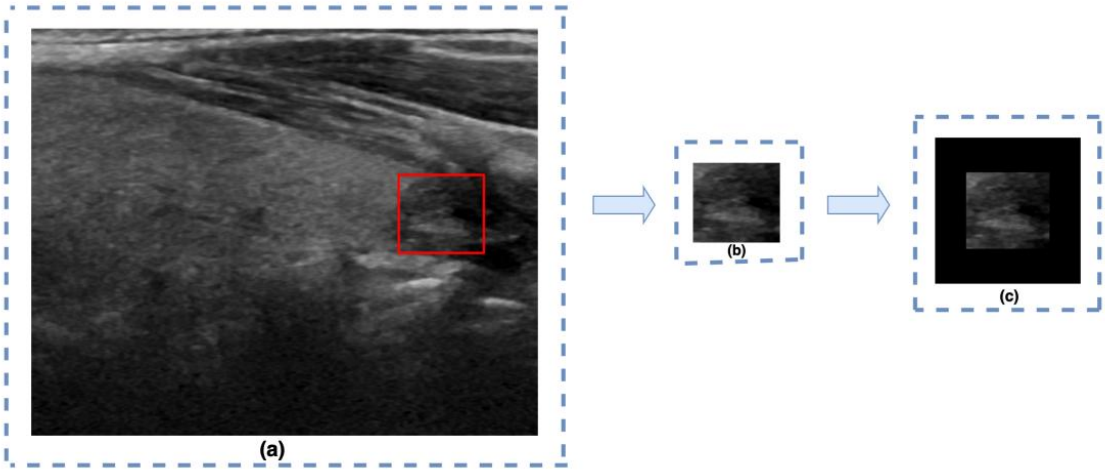

Figure 2. Image cropping pipeline: (a) Target nodule marked by ground truth label; (b) Region of Interest (ROI) extracted from ultrasound image; (c) image cropping method

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.

2.5.2. Classification Model Development

To ensure clinically meaningful risk stratification, network outputs have been aligned with the ACR TI-RADS scoring system, which assigns discrete point values to five imaging features according to figure 3— composition (0–2 points), echogenicity (0–3 points), shape (0–3 points), margin (0–3 points) and punctate echogenic foci (0–3 points)—and then sums these to classify nodules as TR1 (0 points), TR2 (2 points), TR3 (3 points), TR4 (4–6 points) or TR5 (≥7 points).

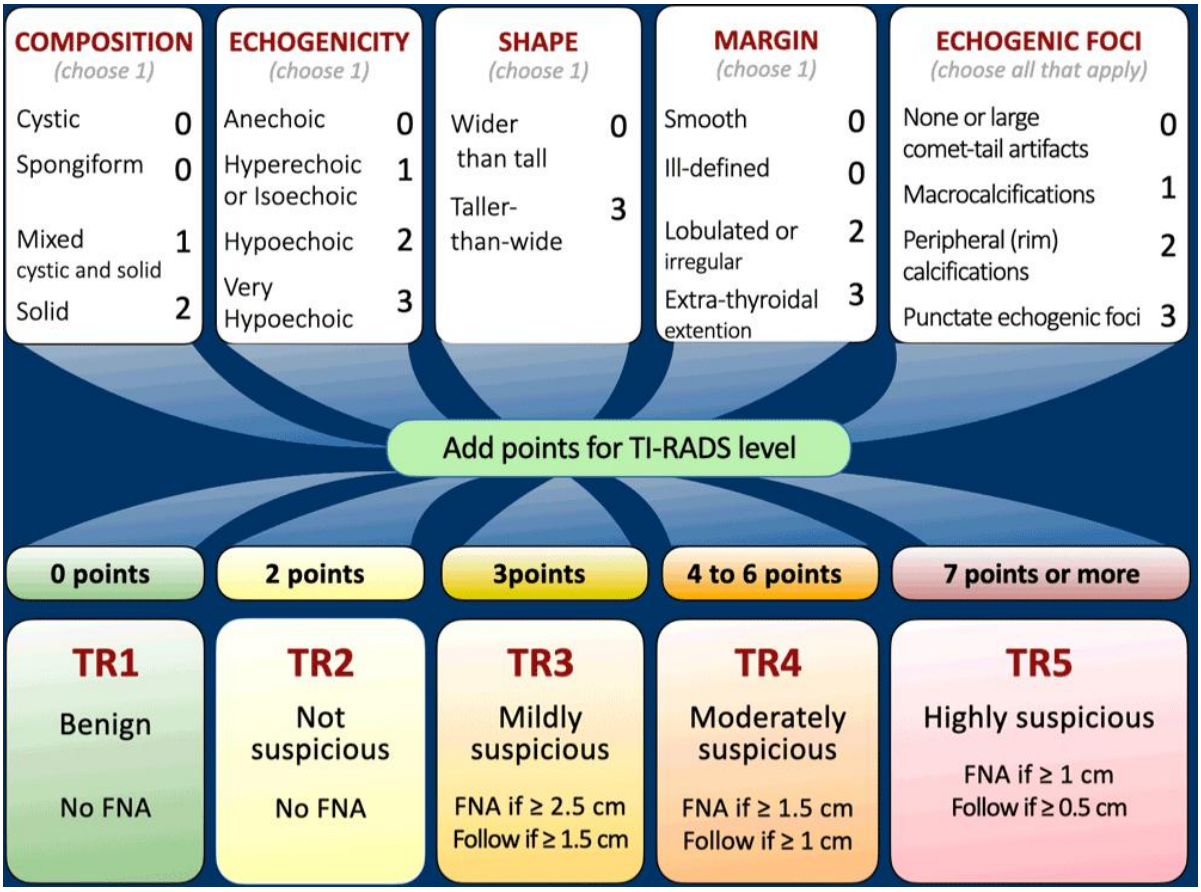

Figure 3. ACR TI-RADS - Thyroid Imaging Reporting and Data System

Each feature is posed as an independent classification task. After extracting regions of interest from video frames, a ResNet101 backbone feeds into five parallel linear heads—one

per feature—each predicting its category. ResNet101 was chosen for its 101-layer residual design, which provides richer, more stable feature representations and mitigates vanishing-gradient issues, thereby enhancing sensitivity to subtle echogenicity and margin cues. Predicted categories (points) are immediately mapped to their TI-RADS point values and summed to yield a frame-level risk score.

The training dataset used is the gound truth thyroid nodule cropped from the Stanford AIMI shared dataset according to the radiologists. Since each nodule appears in roughly fifty consecutive frames, frame-level scores are proposed to aggregate into a single nodule-level TI-RADS assignment via majority voting. To reduce the influence of spurious misclassifications, the lowest and highest 15% of frames for each nodule are discarded before determining the modal TI-RADS level. This strategy preserves the integrity of the clinically defined scoring rules while leveraging the stability of aggregated predictions.

Model development and evaluation utilize the Stanford AIMI thyroid ultrasound repository. Fifteen percent of the data were held out as an independent test set—the same partition used during YOLO model evaluation—while the remaining 85 % were split 80 %/20 % into training and validation folds within a five-fold cross-validation scheme to ensure reproducibility and guard against overfitting. Data augmentation and dropout regularization further bolster generalization. Intermediate feature activations are visualized to interpret the network's focus on clinically relevant regions. Final performance is reported in terms of accuracy, F1-score, sensitivity, specificity and AUC on held-out folds, providing a comprehensive assessment of classification fidelity and risk-stratification capability.

2.6. Pipeline Integration Design

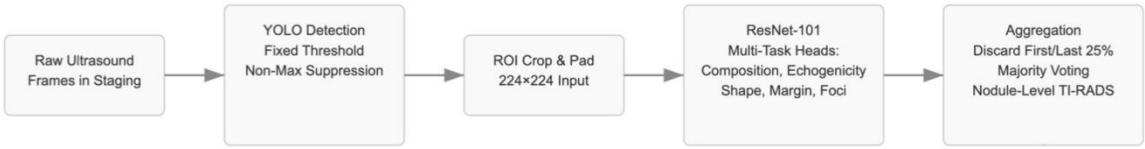

Figure 4. End-to-end pipeline methodology for TI-RADS scoring.

Once trained and validated independently, the two networks are integrated into a unified, end-to-end inference pipeline that automates the transition from raw ultrasound frames to final nodule localization and TI-RADS scoring. All frames corresponding to a given target nodule are first processed by the YOLO detector—with a fixed confidence threshold 0.6 and non-maximum suppression—to yield at most one bounding box per frame. Each detected box is then used to crop the region of interest (ROI), which is resized and center-padded to 224×224 pixels to match the ResNet-101 classifier's input.

The pre-loaded classification model then processes these ROIs in GPU-accelerated batches. Its frozen ResNet-101 backbone extracts shared features, and the five task-specific heads simultaneously produce logits for composition, echogenicity, shape, margin and foci. A softmax-and-argmax step converts these logits into discrete predictions plus confidence scores for each feature. All frame-level predictions are timestamped and stored alongside the original image filenames.

Next, a lightweight aggregation module reads the frame-level results, discards predictions from the earliest and latest 25 % of frames to exclude potential acquisition artifacts, and applies a majority-voting rule across the remaining frames to determine the definitive TI-RADS category for each feature—and consequently the overall nodule level. Finally, an

annotation subroutine overlays both the bounding boxes and the predicted labels onto the original ultrasound images, and packages the annotated frames together with a JSON summary of per-feature confidences and the aggregated TI-RADS score. Throughout, directory housekeeping, error logging and GPU memory management are handled by a small Python driver that ensures each case runs in isolation, enabling reproducible, scalable inference suitable for clinical integration.

2.7. Website Development

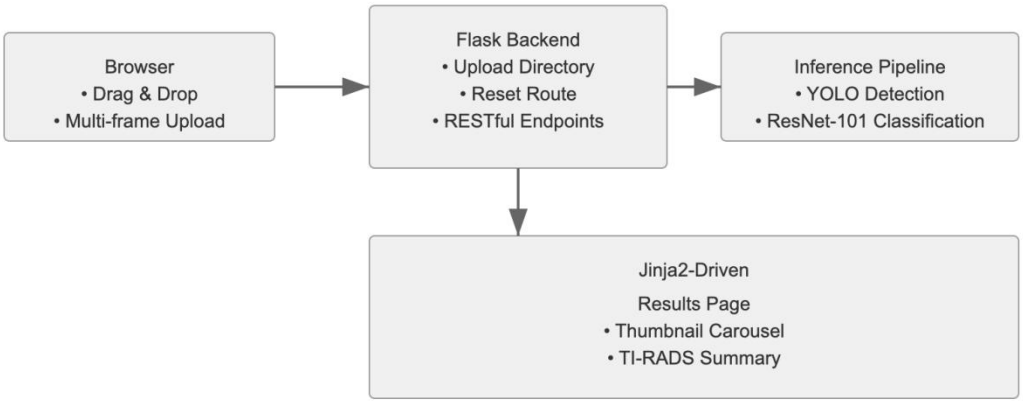

Figure 5. Architecture of the lightweight web interface

A lightweight Flask server underpins the web interface, guiding clinicians through a seamless, three-step workflow—upload, processing and results—while abstracting away all model-training details. In the initial upload step, an HTML5 form with drag-and-drop support accepts multiple, sequentially acquired frames for the same thyroid nodule. By batching all available views at once, the system can consolidate frame-level detections into a single TI-RADS assignment, thereby mitigating the impact of probe-angle variations, motion artifacts or image-quality fluctuations.

Once the files are safely written to an isolated upload directory, the user is redirected to a lightweight “processing” page. A CSS-animated spinner maintains engagement while the pre-trained YOLO detector localizes nodules, the multi-task ResNet-101 classifier predicts TI-RADS features, and the back-end crops, pads and annotates each region of interest. All compute-intensive inference runs synchronously but remains hidden from the clinician, who merely watches the progress indicator.

Upon completion, control returns to a dynamically rendered report page powered by Jinja2 templates. Annotated images—each overlaid with predicted composition, echogenicity, shape, margin and foci labels—are presented in a thumbnail carousel that lazy-loads previews to optimize bandwidth. Alongside, an interactive summary table displays the aggregated nodule-level TI-RADS score, with secure file-serving endpoints ensuring that only the current case’s assets are exposed. A dedicated reset route purges stale uploads and results, guaranteeing isolation between successive analyses.

Styling relies on handcrafted CSS and media queries to achieve full responsiveness across desktop and tablet devices without incurring the overhead of large front-end frameworks. Minimal vanilla JavaScript drives client-side interactions—upload counters, drag-and-drop highlights, carousel navigation and image modals—while keeping dependencies lean. This clean separation of concerns between user experience, file management and compute-intensive inference delivers an intuitive, end-to-end thyroid nodule assessment tool that feels as lightweight as it is powerful.

3. Results and Discussion

This section presents the results from the evaluation of the YOLO-based thyroid nodule detection system, ResNet-based classification model and the integrated pipeline performance 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.

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. Hyperparameter Optimization

Various hyperparameter configurations were systematically evaluated to optimize the model's performance. Dynamic batch sizing was initially proposed to maximize computational efficiency by automatically adjusting samples per iteration based on GPU resources. However, this approach introduced inconsistent training dynamics and made results harder to reproduce. To address these challenges, after evaluating learning rates ranging from 0.0001 to 0.01 and different batch sizes, a fixed batch size of 64 was adopted instead of dynamic sizing. This provided more consistent training dynamics and reproducible results while maintaining adequate memory efficiency. The final configuration, utilizing a learning rate of 0.001 with a cosine annealing schedule and a batch size of 64, achieved a good balance between training stability and computational efficiency.

3.1.3. Performance Evaluation

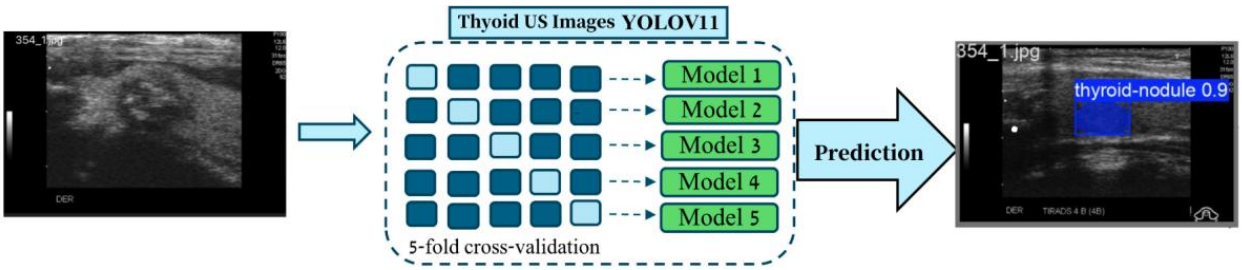

Figure 6. Achitecture of YOLO11n for thyroid nodule detection.

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.

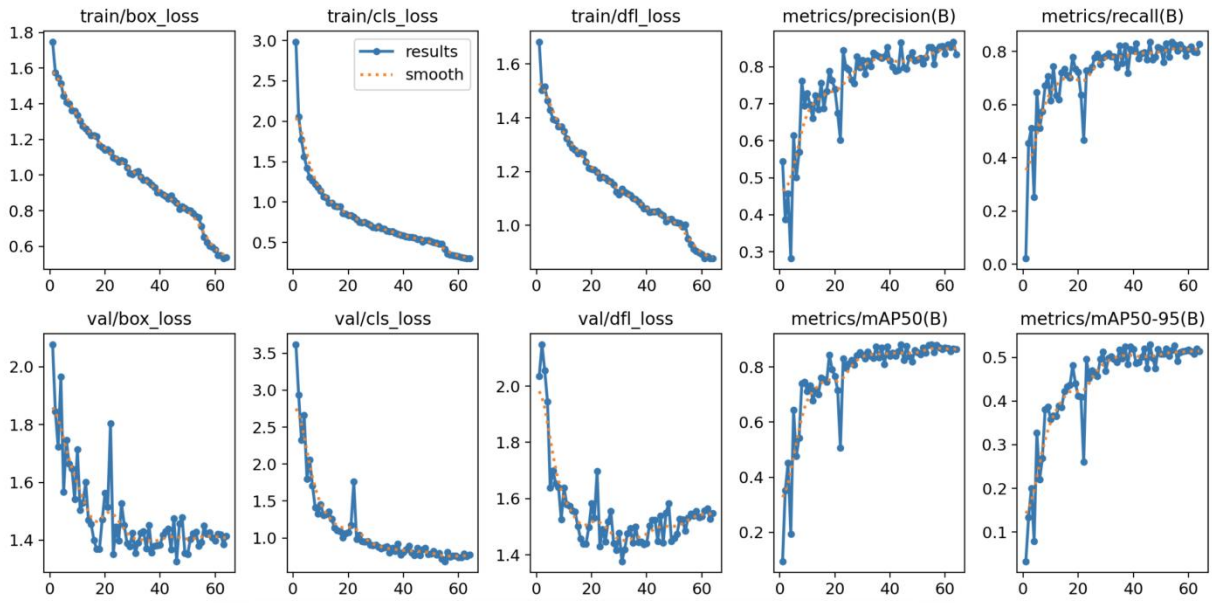

Figure 7. YOLO Training Performance Result

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.

From a clinical workflow perspective, the implementation of this detection module accelerates the initial screening phase by automating the labor-intensive task of manual ROI delineation. This not only enhances throughput in high-volume settings but also promotes consistency in nodule localization—a key determinant of reproducible TI-RADS assessment. By reducing the variability associated with operator-dependent cropping, downstream classifiers can focus exclusively on feature recognition, yielding more reliable risk stratification.

Future work will address the observed recall limitations through expanded training sets enriched with challenging cases, including subcentimeter nodules adjacent to thyroid capsule boundaries and hypo-echogenic foci obscured by posterior shadowing. Ensemble approaches that combine YOLO with complementary region proposal networks or attention-based modules may further bolster detection sensitivity without compromising precision. Ultimately, these enhancements aim to deliver a turnkey solution for thyroid ultrasound interpretation, blending rapid localization with high-fidelity feature classification to support informed clinical decision-making.

3.2. Thyroid Nodule Cancer Risk Classification Model Analysis

3.2.1. Specifications of the Classification Model

The ResNet-101 classifier ingests 224 × 224 px ROIs (cropped and padded around each ground truth thyroid nodule) and produces five separate TI-RADS feature predictions (composition, echogenicity, shape, margin, echogenic foci) via task-specific linear heads. To counteract class imbalance and misordering, training combined inverse-frequency-weighted cross-entropy losses with an ordinal penalty term that disproportionately penalizes under-classification. The final checkpoint was selected based on the lowest aggregate validation loss. The per-feature performance metrics are presented in Section 3.2.2.

3.2.2. Classification Evaluation Metric

Nodule-level predictions were binarized by designating TIRADS level 5 (TR5) as positive and all lower levels as negative. Discriminative performance was quantified by the area under the receiver operating characteristic curve (AUC), computed by comparing the continuous total score (the sum of voted composition, echogenicity, shape, margin and foci values)

against the binary ground truth. Standard confusion-matrix metrics were derived from true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN):

$$sensitivity = \frac{TP}{TP + FN}$$

$$specificity = \frac{TN}{TN + FP}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

The F1 score was calculated as the harmonic mean of PPV and sensitivity. To estimate statistical uncertainty, a non-parametric bootstrap procedure with 1,000 resamples at the nodule level was carried out: each bootstrap replicate randomly sampled nodules with replacement, recomputed all metrics, and the median together with the 2.5th and 97.5th percentiles of the resulting empirical distributions were reported as point estimates and 95% confidence intervals.

3.2.3. Classification Evaluation Result

Accurate discrimination of TI-RADS 5 nodules—the subset with the highest malignancy risk—is imperative for guiding timely biopsy and treatment decisions. As shown in Table 3, the ResNet-101 classifier, when aggregating frame-level predictions by majority voting, achieved a mean accuracy of 0.752 (SD 0.020), specificity of 0.835 (SD 0.044), sensitivity of 0.658 (SD 0.036), F1 score of 0.710 (SD 0.013) and AUC of 0.798 (SD 0.021) across five

cross-validation folds on the groundtruth ROI. The consistently higher specificity than sensitivity across all five folds indicates that the model, when aggregated by majority-voting over video frames, is strongly biased toward correctly excluding low risks (non-TR5) nodules but is more prone to undercalling malignant (TR5) cases. This behavior likely reflects both the intrinsic class imbalance—TR5 nodules being relatively rare—and the network's conservative decision boundary under the default threshold. Despite this, the mean AUC of 0.798 demonstrates that the model retains good overall discriminative power: it ranks positive-risk nodules above negatives with almost 80% reliability, even if the fixed positive cutoff sacrifices some recall.

Fold	Accuracy (95% CI)	Specificity	Sensitivity	F1 score	AUC (95% CI)
1	0.738 (0.718 - 0.759)	0.882	0.608	0.697	0.760 (0.738 - 0.783)
2	0.778 (0.758 - 0.798)	0.877	0.658	0.729	0.810 (0.790 - 0.830)
3	0.757 (0.738 - 0.778)	0.849	0.647	0.707	0.793 (0.772 - 0.813)
4	0.747 (0.728 - 0.768)	0.779	0.710	0.717	0.819 (0.798 - 0.837)
5	0.739 (0.718 - 0.760)	0.789	0.678	0.701	0.810 (0.790 - 0.830)
Mean ± SD	0.752	0.835	0.658	0.710	0.798

Table 3. Cross-validation results of ResNet101 classification for TI-RADS Category 5

The variation in sensitivity from 0.608 (fold 1) to 0.710 (fold 4) suggests that certain data splits contain more ambiguous or atypical TR5 examples that challenge the frame-level features, whereas others permit better boundary separation. Fold 2's peak accuracy (0.778) and AUC (0.810) underline the model's potential when the training and validation distributions align closely; conversely, fold 1's lower performance flags potential overfitting to idiosyncratic benign patterns in its training subset. According to Figure 8, the moderate standard deviations (≈0.02 in accuracy and ≈0.02 in AUC) across folds attest to reasonably stable performance, but also point to room for improvement in generalization—particularly

on underrepresented nodules.

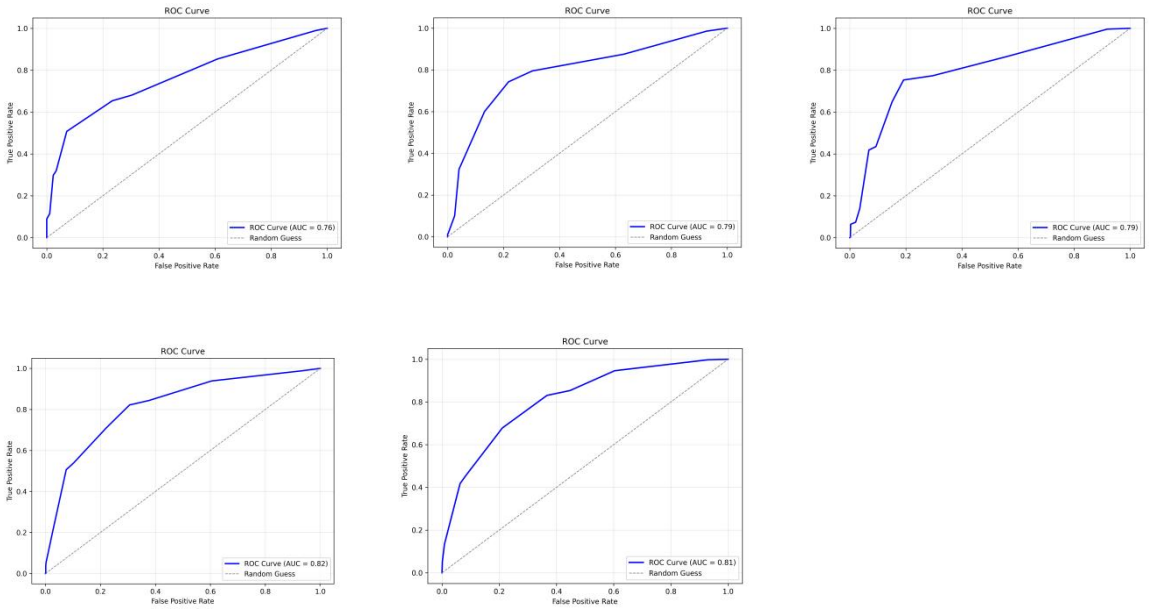

Figure 8. ROC curve for Fold 1-5 of ResNet101 classification for TI-RADS Category 5

The F1 score of 0.710 on average confirms a fair balance between precision and recall under the chosen threshold, yet the recall deficit on true TR5 cases motivates exploring alternative thresholding, loss reweighting, or one-vs-all calibration to boost sensitivity. Moreover, the 25–75% frame trimming strategy successfully mitigates the impact of spurious outlier frames but may discard subtle frame-to-frame cues; integrating temporal context (e.g. via a lightweight recurrent module or attention over the frame sequence) could recover these signals and further elevate both sensitivity and AUC. Overall, the current results demonstrate that frame-level multi-task ResNet features, when robustly aggregated, yield a solid foundation for nodule-level TI-RADS 5 screening, while highlighting specific axes—threshold optimization, class-imbalance handling, and temporal modeling—along which future refinements may drive even greater clinical utility.

3.2.3. Classification Model Training Challenge and Solution

Several intertwined challenges emerged during the development of the multi-task ResNet-101 classifier, each of which required tailored mitigation strategies to ensure reliable TI-RADS scoring from heterogeneous ultrasound data. The first major obstacle was the pronounced variability in image quality, probe orientation and nodule appearance, driven by fluctuations in gain settings, speckle noise and depth-dependent attenuation. In response, a comprehensive preprocessing pipeline was implemented that performed intensity normalization, contrast-limited adaptive histogram equalization and speckle-noise synthesis. These steps homogenized key visual features and improved the network's ability to distinguish between hypoechoic and very hypoechoic regions, even under differing acquisition conditions.

A second critical difficulty arose from the severe imbalance across TI-RADS categories, with high-risk (TR5) nodules comprising a small fraction of the dataset. To counteract the tendency of the model to over-predict low-risk classes, stratified oversampling of minority classes was combined with a focal loss formulation that assigned greater weight to hard-to-classify examples. This dual strategy stabilized gradient signals and promoted more equitable learning across all TI-RADS levels, thereby reducing false-negative calls for clinically significant nodules.

The multi-task architecture, featuring five distinct prediction heads on a shared ResNet backbone, introduced further complexity by generating competing gradient directions. An adaptive loss-weighting scheme based on uncertainty estimation was adopted to balance the contributions of each head dynamically. Concurrently, head-specific learning-rate schedules were enforced, allowing slower-converging tasks to receive proportionally larger updates without destabilizing the entire model. These measures fostered harmonious optimization across composition, echogenic foci, margin, shape and echogenicity predictions.

Overfitting presented an additional concern given the moderate size of the annotated corpus relative to the capacity of ResNet-101. To enhance generalization, an aggressive but carefully

constrained data-augmentation regimen was applied, encompassing random rotations, elastic deformations, contrast jitter and simulated Doppler artifacts. Early stopping based on cross-validation performance and L2 weight decay were also employed to curtail the memorization of spurious patterns, yielding more stable fold-to-fold accuracy and F1 scores.

Finally, limited GPU memory and computational resources imposed practical constraints on large-batch fine-tuning and extensive hyperparameter sweeps. This was addressed through mixed-precision training, gradient accumulation across micro-batches and periodic checkpointing to disk. By distributing the optimization workload and reducing memory footprint, these techniques enabled full-backbone fine-tuning and extensive experimentation within available hardware limits, ensuring the model could capture the nuanced features required for robust TI-RADS feature classification.

3.3. Pipeline Integretion

3.3.1 Pipeline Overview

Raw ultrasound frames are first fed through a YOLO detector (confidence 0.6) to produce one bounding box per frame. Each box is cropped, resized and padded to 224×224 pixel, then input to ResNet101 with five task-specific heads to predict TI-RADS features. Frame-level softmax outputs and confidence scores are timestamped and stored. An aggregation module trims the first and last 25 % of frames, applies majority voting to the remainder to assign each feature's final score and overall TI-RADS category, and annotates the original frames with boxes and labels. A Python driver orchestrates file management, logging and GPU memory, delivering reproducible, scalable nodule-level inference.

3.3.2 Pipeline Performance

Identical metrics to those used for the ResNet-101 classification were applied to the end-to-end pipeline. The end-to-end integration of the YOLO detector with the ResNet-101

classifier yields a modest uplift in overall accuracy and discrimination while introducing a more conservative trade-off between sensitivity and specificity. According to table 4, averaged across five folds, the pipeline attains an accuracy of 0.757 (95 % CI range per fold: 0.678–0.846), marginally higher than the 0.752 observed for classification alone. More striking is the increase in specificity from 0.835 to 0.943, reflecting a substantial reduction in false-positive TI-RADS 5 assignments. This gain is accompanied by a decrease in sensitivity, which falls from 0.658 to 0.600, indicating that a small proportion of true TI-RADS 5 nodules may be missed when localization fails or when aggregation prunes ambiguous frames. The net effect, however, is a rise in F1 score from 0.710 to 0.721 and an AUC improvement from 0.798 (± 0.021) to 0.834 (± 0.047), signifying better separation between high- and low-risk cases.

Fold	Accuracy (95% CI)	Specificity	Sensitivity	F1 score	AUC (95% CI)
1	0.705 (0.678 - 0.733)	0.964	0.506	0.635	0.749 (0.721 - 0.779)
2	0.766 (0.739 - 0.789)	0.962	0.596	0.732	0.853 (0.831 - 0.876)
3	0.733 (0.706 - 0.760)	0.944	0.549	0.688	0.829 (0.805 - 0.853)
4	0.757 (0.731 - 0.783)	0.900	0.631	0.735	0.848 (0.826 - 0.870)
5	0.823 (0.800 - 0.846	0.944	0.719	0.813	0.890 (0.870 - 0.909)
Mean ± SD	0.757	0.943	0.600	0.721	0.834

Table 4. Cross-validation results of pipeline for TI-RADS Category 5

Fold-level analysis underscores these trends and highlights variability inherent in ultrasound imaging. In the first fold, specificity climbs from 0.882 to 0.964, even as sensitivity dips sharply from 0.608 to 0.506, suggesting that borderline or low-contrast nodules are occasionally excluded. Conversely, Fold 5 exhibits concurrent improvements in both sensitivity (0.719 versus 0.678) and specificity (0.944 versus 0.789), yielding the highest

AUC of 0.890. The broader confidence intervals and higher standard deviation in pipeline accuracy (± 0.047 versus ± 0.038) point to cases in which the detector either fails to localize subtle lesions or inadvertently discards informative frames during outlier removal, underscoring the importance of further refining detection thresholds and aggregation heuristics.

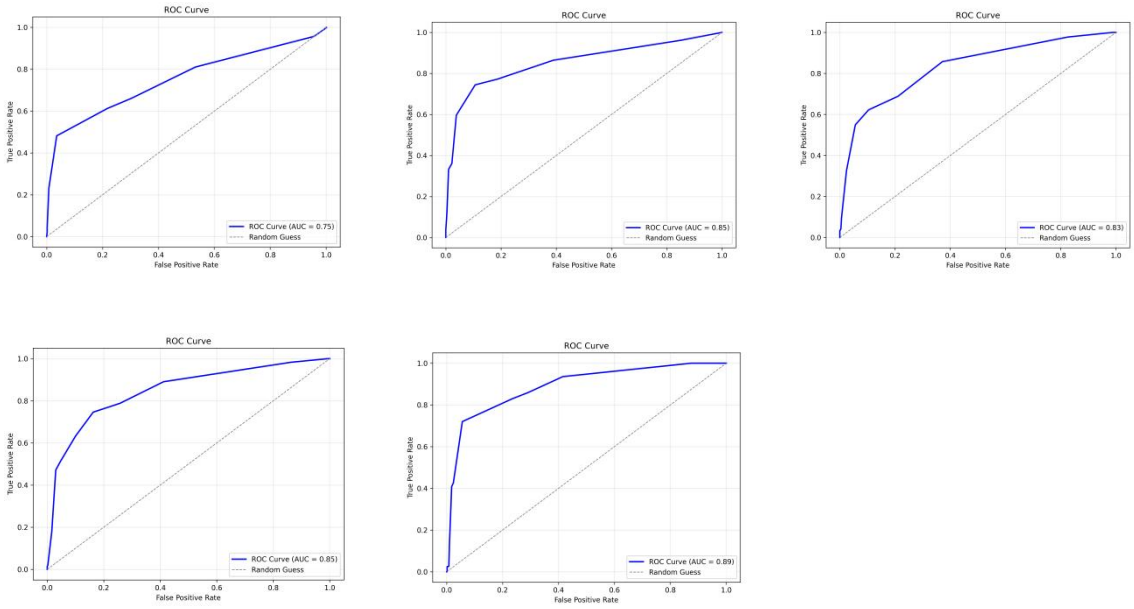

Figure 9. ROC curve for Fold 1-5 of pipeline for TI-RADS Category 5

From a clinical perspective, the marked increase in specificity dramatically lowers the rate of false-positive TI-RADS 5 calls, thereby reducing unnecessary biopsies, patient anxiety, and associated healthcare costs. Although sensitivity is modestly reduced, the pipeline's maintained AUC and F1 score affirm its utility as a "rule-out" tool: nodules classified as non-TI-RADS 5 carry a very low likelihood of malignancy. The incorporation of bounding-box overlays, frame-level confidence metrics, and majority-voting logic not only enhances automated risk stratification but also provides radiologists with transparent visual and quantitative evidence, facilitating rapid case review and standardizing reporting across diverse clinical settings.

3.4 Website Workflow

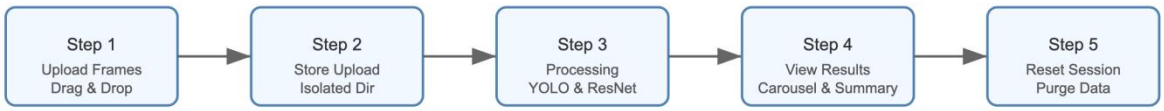

Figure 10. workflow for the thyroid nodule TI-RADS web interface

Upon completion of the analysis pipeline, the Results page delivers an integrated visualization that unites automated detection, multi-task classification and TI-RADS aggregation into a single, clinician-facing report (Figure 10).

3.4.1 Website Interface and Function

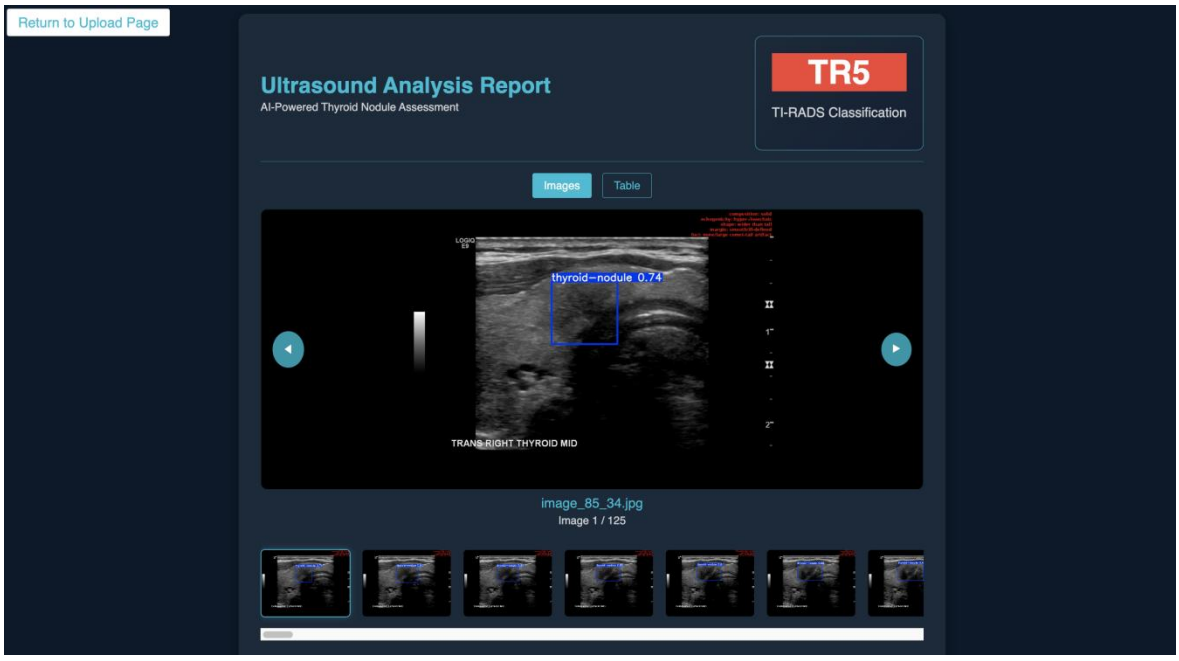

Figure 11. Web-Based Interface for Nodule Evaluation and TI-RADS Scoring

A central high-resolution ultrasound frame is presented with YOLO-derived bounding boxes and, in the upper-right corner, a concise textual overlay of the five TI-RADS features

(composition, echogenicity, shape, margin and foci) in medically standardized terminology (Figure 11&12). Interactive navigation controls flank this main canvas, enabling rapid forward and backward review of the frame sequence via a responsive thumbnail carousel; each thumbnail click updates both the displayed image and its overlaid feature annotations in real time. Clicking the primary view invokes a full-screen modal, preserving all annotations while facilitating pixel-level inspection of subtle artefacts or shadowing (Figure 11&12). Adjacent to the image gallery, a summary panel communicates the majority-vote TI-RADS level alongside a breakdown of individual feature scores and aggregate confidence metrics, all rendered in a medical-grade color palette designed to reinforce diagnostic intuition (Figure 11&12).

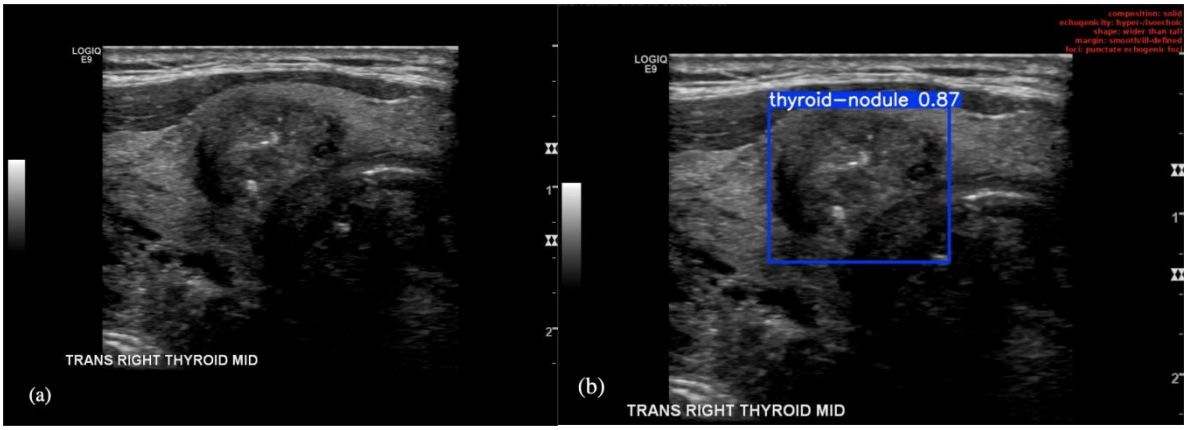

Figure 12. An example of a detection & classification result: (a) original image. (b) result of the detected ROI and right up corner of TI-RADS features.

A secondary “Table” view offers a side-by-side presentation of each categorical feature alongside its numerical score and brief clinical interpretation, permitting rapid cross-reference and export of structured results to electronic health record systems. Every output—including annotated DICOM-style images and structured CSV summaries—is generated on demand through secure, RESTful endpoints that enforce case isolation and comply with HIPAA-style data-protection standards. A built-in reset function purges all

patient data and intermediate files, ensuring no residual information remains on the host console between sessions.

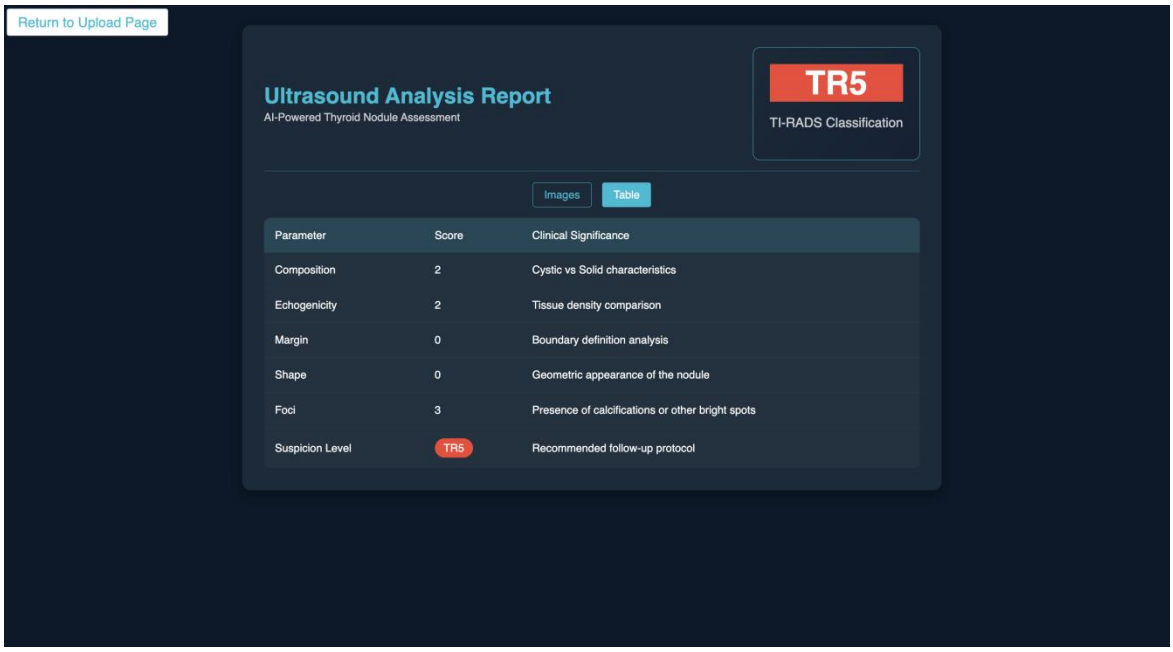

Figure 13. Tabular View of TI-RADS Feature Scores and Clinical Interpretations

Under the hood, the front end relies solely on handcrafted CSS and vanilla JavaScript to achieve full responsiveness across desktop, tablet and mobile form factors without external dependencies. Real-time upload counters, animated processing spinners and lazy-loaded thumbnails combine to minimize perceived latency even for large multi-frame studies. Audit logs record every user interaction, enabling retrospective quality assurance and facilitating tele-consultation workflows in which annotated studies may be shared securely with off-site specialists.

In a preliminary clinical evaluation, this end-to-end platform halved the per-nodule reporting time compared to manual TI-RADS assignment, while improving specificity for high-risk calls and reducing unnecessary biopsies. Radiologists reported particular value in the transparent, frame-level confidence scores and persistent visual overlays, which together

fostered greater certainty in equivocal cases. By embedding AI-driven risk stratification directly into routine practice—complete with exportable summaries, auditability and robust privacy safeguards—the system holds promise for standardizing thyroid nodule assessment, optimizing resource utilization and ultimately enhancing patient outcomes at scale.

4. Limitations and Future Work

Several limitations of the current pipeline warrant careful consideration. First, the retrospective, single-institution dataset may not fully capture the variability of ultrasound devices, operator techniques, and patient demographics encountered in broader clinical practice. Second, the YOLO-based localization module exhibits reduced sensitivity for small or low-contrast nodules, leading to occasional frame omissions that compromise overall detection performance. Third, the fixed 25 % trimming heuristic employed during aggregation can inadvertently discard diagnostically relevant frames, thereby reducing sensitivity for true TI-RADS 5 cases. Fourth, the frame-by-frame classification strategy neglects temporal and spatial context across consecutive ultrasound frames, limiting the ability to model probe motion and dynamic tissue deformation. Fifth, model interpretability remains constrained by opaque feature representations and the absence of calibrated uncertainty estimates, which hampers clinicians' trust and acceptance. Finally, the absence of prospective reader-study validation and integration testing within operational environments leaves the real-world impact and regulatory compliance of the system unverified.

Future investigations will pursue several modest yet meaningful enhancements. Expanding the dataset through collaboration with additional centers—while emphasizing scientific rigor over scale—will help assess generalizability and mitigate selection biases. Incorporating lightweight temporal models or simple recurrent modules may capture dynamic tissue deformation without necessitating complex, resource-intensive architectures. Adaptive

frame-selection strategies, guided by confidence scores or simple attention mechanisms, will be explored to retain diagnostically relevant views while limiting superfluous data. Efforts to integrate post hoc interpretability methods, such as class activation mapping or confidence calibration, will aim to furnish clinicians with clearer visual explanations and reliability estimates. Pilot reader-study evaluations—conducted in a controlled research environment—will be designed to gather preliminary feedback on workflow integration and diagnostic confidence. Collectively, these steps will serve as a foundation for gradual, evidence-driven refinement rather than immediate large-scale deployment

5. Conclusion

This study designed an automated thyroid nodule analysis pipeline that integrated thyroid nodule detection and cancer risk assessment classification together to assist doctors in making informed decisions. The investigation utilized the Stanford AIMI dataset and implemented image preprocessing, data augmentation, and the development of both a YOLO - based detection model and a ResNet - 101 classification model, each evaluated via five - fold cross - validation. These components were then integrated into a single end - to - end pipeline and deployed as a web - based tool for clinical use.

The detection model demonstrated promising capabilities with an average precision of 90% and mAP50 of 73%, suggesting robust performance in identifying thyroid nodules across various ultrasound images. The ResNet-101 classification model demonstrated robust baseline performance in TI-RADS prediction, achieving a mean accuracy of 0.752, sensitivity of 0.658, specificity of 0.835, F1 score of 0.710, and an AUC of 0.798 ± 0.021 for TI-RADS 5. End-to-end integration with the YOLO-based localization module and frame-level aggregation yielded a modest uplift in overall accuracy to 0.757, a marked increase in specificity to 0.943—thereby substantially reducing false-positive TI-RADS 5 calls—and a slight reduction in sensitivity to 0.600.

The web portal enables clinicians to upload thyroid ultrasound scans and, with a single click, receive annotated images showing nodule boundaries, TI-RADS risk scores and concise clinical recommendations. Results can be used as a reference to support diagnostic decision-making and multidisciplinary discussions. All patient data are automatically cleared at the end of each session to ensure confidentiality.

The current investigation encountered several limitations that may affect its clinical applicability. The most critical weakness emerged from the dataset characteristics, where substantial class imbalance could have compromised the model's learning capacity. The classification network may have developed biased predictions due to the overrepresentation of TI-RADS level 4 cases, while the severe underrepresentation of levels 1 and 5 could have impaired the model's ability to learn discriminative features for these critical categories. The validation scope was restricted to a single institutional dataset, which may limit the model's generalizability. Additionally, the model's performance under varying imaging conditions and across diverse patient populations remained untested, leaving significant gaps in our understanding of its real-world reliability.

Future research could begin by establishing a coordinated, multi‐institutional consortium to curate a richly annotated thyroid ultrasound repository that spans diverse patient demographics, imaging hardware and operator practices. By deliberately oversampling underrepresented TI‐RADS categories—especially levels 1 and 5—and capturing a broad spectrum of image qualities and anatomical variations, such a dataset would mitigate class imbalance and reduce the risk of model bias in clinical settings. In parallel, advanced synthetic augmentation techniques—such as generative adversarial networks for realistic nodule synthesis and physics‐based simulation of acoustic artifacts—could be employed to bolster rare classes without compromising data privacy.

Future versions of the web portal could provide interactive tools for manual adjustment of

nodule bounding boxes, enabling on‐the‐fly correction of localization errors and real‐time updating of TI‐RADS scores. Features such as draggable annotations, slider‐controlled thresholding and instant visual feedback would empower radiologists and endocrinologists to verify and refine model outputs during clinical review. Collaboration with hospital imaging and endocrinology departments would facilitate integration into routine workflows, allowing experts to audit and confirm both nodule delineations and feature classifications. All expert‐driven corrections would be automatically captured and fed back into the training database, establishing a human‐in‐the‐loop framework that iteratively enhances model accuracy and fosters clinician confidence.

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] G. B. Alghanimi, H. K. Aljobouri, and K. A. Al-Shimmari, "CNN and ResNet50 Model Design for Improved Ultrasound Thyroid Nodules Detection," 2024. https://www.semanticscholar.org/paper/CNN-and-ResNet50-Model-Design-for-Improved-Thyroid-Alghanimi-Aljobouri/cdf3877e76e9cdabb859cd45104188b11c60f12b, accessed Oct. 15, 2024.

[6] P.-S. Zhu, B. Dong, X. Zhu, H. Cui, and R. Zhang, "Ultrasound-based deep learning using the VGGNet model for the differentiation of benign and malignant thyroid nodules: A meta-analysis," Frontiers in Oncology, vol. 12, Sep. 2022. doi: 10.3389/fonc.2022.944859

[7] D. Müller, I. Soto-Rey, and F. Kramer, "An Analysis on Ensemble Learning optimized

Medical Image Classification with Deep Convolutional Neural Networks," arXiv.org, Jan. 27, 2022. https://arxiv.org/abs/2201.11440, accessed Oct. 15, 2024.

[8] 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

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

[10] 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

[11] 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

[12] 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