

# Artificial Intelligence–Assisted Ultrasound Interpretation Enhances Diagnostic Performance Among Trainees in Thyroid Imaging

Fumihiko Furuya

[ffuruya@fmu.ac.jp](mailto:ffuruya@fmu.ac.jp)

Fukushima Medical University

Keiichi Nakano

Fukushima Medical University

Yoshiko Matsumoto

Fukushima Medical University

Koki Shio

Fukushima Medical University

Satoshi Suzuki

Fukushima Medical University

Satoru Suzuki

Fukushima Medical University

Hiroki Shimura

Fukushima Medical University

Hidetoshi Ando

Graduate School of University of Yamanashi

---

## Research Article

### Keywords:

**Posted Date:** April 21st, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-9329400/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Abstract

## Background:

Accurate interpretation of thyroid ultrasound requires substantial experience, and diagnostic performance among trainees remains variable. Artificial intelligence (AI) has shown promise in medical image analysis; however, its role as an educational support tool for trainees has not been fully established. This study aimed to evaluate whether an AI-assisted system could improve trainee performance in thyroid ultrasound interpretation.

## Methods:

We developed an AI-based diagnostic support system using a Mask R-CNN framework trained on pathology-confirmed thyroid ultrasound images. The model performed segmentation and classification of five lesion categories. A reader study was conducted involving five expert physicians and five trainees. Trainees evaluated ultrasound images under non-assisted and AI-assisted conditions. Diagnostic performance was assessed using sensitivity, specificity, precision, and F1 score.

## Results:

The AI model demonstrated high segmentation accuracy for several structures, including vessels and thyroid parenchyma. However, its ability to distinguish benign from malignant tumors remained limited. Despite this, AI assistance improved trainee performance. Median precision increased from 0.53 to 0.63, and F1 score improved from 0.53 to 0.61. Specificity increased from 0.22 to 0.50, while sensitivity remained comparable. These findings indicate that AI support primarily reduced false-positive interpretations and improved diagnostic consistency among trainees.

## Conclusions:

AI-assisted ultrasound interpretation improved the diagnostic performance of trainees, particularly by enhancing precision and reducing false-positive findings. Although the AI model alone was insufficient for independent clinical use, it demonstrated potential as an educational support tool. These findings suggest that AI may contribute to more consistent and effective training in thyroid ultrasound interpretation.

## Introduction

Thyroid ultrasound is widely used for the evaluation of thyroid morphology and the characterization of nodular lesions. Accurate interpretation of ultrasound findings is essential for appropriate clinical decision-making, including the indication for fine-needle aspiration cytology. However, ultrasound interpretation is highly operator-dependent and requires substantial training and experience, leading to variability in diagnostic performance among clinicians.

In large-scale examination settings, such as those implemented following the Chernobyl nuclear disaster, there is an increased demand for rapid and consistent image interpretation [1, 2]. In such situations, the limited

availability of experienced specialists poses a significant challenge, and the need for effective training strategies for novice clinicians becomes increasingly important.

Artificial intelligence (AI), particularly convolutional neural networks (CNNs), has shown promising performance in medical image analysis, including classification, object detection, and segmentation tasks. In thyroid ultrasound, AI has been primarily investigated as a tool for automated diagnosis, focusing on distinguishing between benign and malignant lesions [3, 4].

However, despite these advances, the role of AI as an educational support tool for trainees has not been fully established. Given that clinical decision-making still requires human judgment, it is important to explore how AI can be integrated into training to enhance diagnostic performance and consistency among less experienced clinicians.

In this study, we developed an AI-based diagnostic support system using a Mask R-CNN framework trained on pathology-confirmed thyroid ultrasound images. We aimed to evaluate not only its diagnostic performance but also its impact on trainee interpretation of thyroid ultrasound images. Specifically, we investigated whether AI assistance could improve diagnostic performance and reduce variability among novice trainees.

## Materials and Methods

### *Development of the AI Diagnostic Support Tool*

Preoperative ultrasound images from patients with surgically treated and histopathologically confirmed thyroid lesions were retrospectively analyzed. Ground truth annotations were created for five classes: thyroid parenchyma, lymph node, vessel, benign tumor (including follicular adenoma, adenomatous nodule, and cyst), and malignant tumor. All cases in the malignant tumor category were pathologically diagnosed as papillary thyroid carcinoma. Each ultrasound image could include multiple classes within the same frame. The dataset consisted of 1,206 images divided into training ( $n = 700$ ), validation ( $n = 254$ ), and test ( $n = 252$ ) sets, derived from 340 patients. To prevent data leakage, all images from a single patient were assigned exclusively to one subset.

To mitigate class imbalance, data augmentation was applied during training, including horizontal flipping. These procedures approximately doubled the number of minority-class samples. Additionally, five-fold cross-validation was performed using the same hyperparameter settings to verify model robustness. Cross-validation results demonstrated consistent segmentation performance across folds (mean mAP =  $0.78 \pm 0.03$ ; mean IoU =  $0.81 \pm 0.02$ ), confirming that augmentation and balanced sampling effectively reduced bias toward majority classes.

### *Ethical Approval*

This study was approved by the Ethics Committees of Fukushima Medical University (approval number: 2022192) and University of Yamanashi (approval number: 2740). Written informed consent was obtained from all participants. The ultrasound image data were accessed for research purposes from 01/05/2024 to 10/10/2025. All procedures performed in this study were in accordance with the ethical standards of the institutional research committees and with the Declaration of Helsinki.

### *Participants and Diagnostic Accuracy Evaluation*

Five board-certified thyroid imaging expert doctors with over 10 years of experience and five medical trainees from Fukushima Medical University who had passed the Japanese national Objective Structured Clinical Examination and Computer-Based Testing were recruited. They were presented with 35 ultrasound images from 9 cases with confirmed pathology (including 29 malignant and 18 benign lesions). The AI's predictions were matched with ground truth labels based on maximum Intersection over Union (IoU). The prediction was counted as correct if the class label and region matched.

To evaluate the effect of AI support on diagnostic performance, all participants reviewed the same set of thyroid ultrasound images under two conditions. In the first round ("non-assisted" session), trainees interpreted the original ultrasound images without any AI-generated information and annotated lesions based solely on their own judgment. In the second round ("AI-assisted" session), the same trainees evaluated identical images with Mask R-CNN-generated segmentation overlays, class labels (benign or malignant), and confidence scores displayed on the screen. The AI output highlighted predicted lesion areas and showed the model's confidence for each classification. Participants could view these AI-generated suggestions but were not allowed to modify them. All ultrasound images were obtained from patients who subsequently underwent thyroid surgery, and final pathological diagnoses were confirmed postoperatively.

### *Statistical Analysis*

Group comparisons were performed using the Mann–Whitney U test. Statistical analyses were performed using STATA software (version 14.2; Stata Corp., College Station, TX, USA). Statistical significance was defined as  $p < 0.05$ .

## **Results**

### *Performance of the AI Tool*

The AI achieved promising segmentation accuracy on the test dataset. Table 1 summarizes the distribution of ultrasound images used for the development and evaluation of the AI diagnostic model. The dataset was categorized into five anatomical/pathological classes: parenchyma, lymph node, vessel, benign tumor (including follicular adenoma, adenomatous nodule, and cyst), and malignant tumor.

Images were randomly allocated into training, validation, and test sets. In the test dataset, the model detected 210 thyroid parenchyma regions (92.3% detection rate), 2 lymph node regions (20.0%), 130 vessel regions (82.8%), 101 benign tumor regions (82.8%), and 65 malignant tumor regions (69.1%). The Intersection over Union (IoU) values for the detected regions were 0.80 (95% CI: 0.76–0.82) for thyroid parenchyma, 0.89 (95% CI: 0.88–0.90) for lymph nodes, 0.89 (95% CI: 0.88–0.90) for vessels, 0.80 (95% CI: 0.79–0.84) for benign tumors, and 0.81 (95% CI: 0.76–0.82) for malignant tumors, indicating high localization accuracy in all classes.

Due to the considerable imbalance in lesion type distribution, a confusion matrix was constructed to comprehensively evaluate the classification performance of the model across all lesion categories. As shown in Figure 1, the model demonstrated high sensitivity for vessel lesions (95.1%) and parenchymal structures (66.2%), but frequently misclassified benign and malignant tumors. Notably, 47.0% of pathologically confirmed malignant tumors were correctly identified as malignant tumors, while an equal proportion (47.0%) were misclassified as benign.

## *Comparison with Experts and Trainees*

To evaluate the diagnostic utility of the AI model developed in this study, we assessed its ability to differentiate between benign and malignant thyroid tumors. The model's diagnostic performance was compared with that of expert doctors and trainees. Furthermore, we investigated the AI's effectiveness as a diagnostic support tool for trainees. For this analysis, an additional set of 47 ultrasound images was used, consisting of 18 benign tumors and 29 malignant tumors (all histologically confirmed). Detailed IoU values for each class are shown in Table 2. The AI demonstrated the highest IoU for vessel regions (0.90, 95% CI: 0.80–0.91), followed by malignant tumors (0.83, 95% CI: 0.67–0.93), benign tumors (0.81, 95% CI: 0.75–0.86), and thyroid parenchyma (0.62, 95% CI: 0.48–0.67). These results indicate particularly accurate segmentation of vascular and tumor structures, which are clinically relevant for diagnosis.

The diagnostic performance of trainees was compared between the non-assisted and AI-assisted sessions and further contrasted with that of expert physicians. The performance metrics are summarized in Table 3 and illustrated in Figure 2. Without AI assistance, trainees demonstrated relatively low diagnostic accuracy, with a median precision of 0.53 and F1 score of 0.53. When supported by AI-generated segmentation overlays, class predictions, and confidence scores, their performance modestly improved. Median precision increased from 0.53 to 0.63, and the F1 score from 0.53 to 0.61. Specificity also improved from 0.22 to 0.50, while sensitivity remained comparable (0.55 vs. 0.52). These results indicate that AI assistance mainly enhanced precision and specificity, reducing false-positive interpretations by trainees. However, statistically significant differences ( $p < 0.05$ ) in sensitivity and F1 score remained between AI-assisted trainees and experts, suggesting that the AI tool only partially closed the performance gap.

## **Discussion**

In this study, we developed an AI-assisted diagnostic support system for thyroid ultrasound interpretation and evaluated its impact on trainee performance. The results demonstrated that AI assistance modestly improved diagnostic performance among trainees, particularly in terms of precision and specificity. These improvements were primarily driven by a reduction in false-positive interpretations, suggesting that AI support may enhance the consistency of diagnostic decision-making among less experienced clinicians.

Although the AI model achieved high segmentation performance for several anatomical structures, including vessels and thyroid parenchyma, its classification performance for distinguishing benign from malignant tumors remained limited. As shown in the confusion matrix, nearly half of malignant tumors were misclassified as benign, indicating that the standalone diagnostic performance of the model is insufficient for independent clinical use. However, despite these limitations, the AI system still contributed to improved trainee performance, highlighting its potential value as a supportive educational tool rather than a replacement for expert judgment.

These findings are consistent with the concept that AI can augment, rather than replace, human expertise. Clinical interpretation of ultrasound images requires integration of multiple sources of information, including imaging features, clinical context, and experience-based judgment. In this setting, AI-generated segmentation overlays and probability-based outputs may serve as visual and cognitive aids, helping trainees to better recognize lesion characteristics and avoid overinterpretation.

Previous studies have primarily focused on improving diagnostic accuracy using AI-based classification systems, such as AI-integrated Thyroid Imaging Reporting and Data Systems (AI-TIRADS), which demonstrated improved diagnostic accuracy and interobserver agreement [5]. In contrast, our approach emphasizes pixel-level segmentation combined with classification, providing intuitive visual feedback that may be particularly beneficial in educational settings. Similar findings have been reported in breast ultrasound studies, where AI assistance improved diagnostic performance, especially among less experienced clinicians. Together, these studies support the role of AI as a complementary tool that enhances, but does not replace, clinical expertise.

From an educational perspective, our results suggest that AI-assisted systems may help standardize image interpretation and reduce variability among trainees. In environments where access to expert supervision is limited, such tools could support more efficient learning and improve diagnostic consistency. Importantly, rather than eliminating the need for training, the introduction of AI may redefine the learning process by requiring clinicians to develop skills in interpreting and appropriately utilizing AI-generated outputs.

Several limitations should be acknowledged. First, the dataset size was relatively small and derived from a single institution, which may limit the generalizability of the findings. Second, class imbalance, particularly for lymph nodes, affected model performance despite the use of data augmentation and cross-validation. Third, although AI assistance improved trainee performance, it did not achieve expert-level accuracy, indicating that further refinement of the model and its integration into clinical workflows is necessary.

In addition, future studies should evaluate the impact of AI-assisted training in larger and more diverse cohorts, as well as explore user experience and learning outcomes over time. Improvements in model interpretability, confidence calibration, and user interface design may further enhance the effectiveness of AI as an educational support tool.

In conclusion, our AI-assisted diagnostic system improved the diagnostic performance of trainees in thyroid ultrasound interpretation, particularly by enhancing precision and reducing false-positive findings. While the model is not sufficient as a standalone diagnostic tool, it shows promise as an educational aid that may support more consistent and efficient training in thyroid ultrasound interpretation.

## **Declarations**

### **Acknowledgments**

We sincerely thank the trainees and expert physicians who participated in this study for their valuable contributions.

Ethics approval and consent to participate:

This study was approved by the Ethics Committees of Fukushima Medical University (approval number: 2022192) and University of Yamanashi (approval number: 2740). All procedures were performed in accordance with the Declaration of Helsinki.

Consent to participate:

Written informed consent was obtained from all participants.

Consent for publication:

Not applicable.

Competing interests:

The authors declare that they have no competing interests.

**Data Availability Statement**

The datasets used or analyzed during the current study are available from the corresponding author on reasonable request.

**Author contributions**

FF conceived and designed the study. FF and KN performed data analysis. KS, YM, and SS (Satoshi) acquired the ultrasound images. SS (Satoru), HS, HA contributed to study supervision and interpretation of the data. All authors contributed to manuscript drafting and revision and approved the final version of the manuscript.

**Funding**

This study was supported by commissioned research funding from FREI and the Environmental Health Research Program on Radiation Effects of the Ministry of the Environment, Japan.

**References**

1. Kazakov VS, Demidchik EP, Astakhova LN. Thyroid cancer after Chernobyl. *Nature*. 1992;359(6390):21–21.
2. Zablotska LB, Ron E, Rozhko AV, Hatch M, Polyanskaya ON, Brenner AV, Lubin J, Romanov GN, McConnell RJ, O’Kane P, et al. Thyroid cancer risk in Belarus among children and adolescents exposed to radioiodine after the Chornobyl accident. *Br J Cancer*. 2011;104(1):181–7.
3. He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. *IEEE Trans Pattern Anal Mach Intell*. 2020;42(2):386–97.
4. Cai Z, Vasconcelos N. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *IEEE Trans Pattern Anal Mach Intell*. 2021;43(5):1483–98.
5. Li X, Zhang S, Zhang Q, Wei X, Pan Y, Zhao J, Xin X, Qin C, Wang X, Li J, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet Oncol*. 2019;20(2):193–201.

**Tables**

Table1. Dataset composition used for AI model development and evaluation.

IoU: Intersection over Union. mAP: mean Average Precision. IoU values represent the median for each class; values in parentheses indicate the 95% confidence intervals (CI). mAP was calculated using a threshold of IoU ≥ 0.5.

|                      | <b>parenchyma</b><br><b>(n=1,077)</b> | <b>lymph node</b><br><b>(n=110)</b> | <b>vessel</b><br><b>(n=790)</b> | <b>benign tumor</b><br><b>(n=1,036)</b> | <b>malignant tumor</b><br><b>(n=549)</b> |
|----------------------|---------------------------------------|-------------------------------------|---------------------------------|-----------------------------------------|------------------------------------------|
| <b>training</b>      | 646                                   | 71                                  | 474                             | 731                                     | 287                                      |
| <b>validation</b>    | 205                                   | 29                                  | 159                             | 131                                     | 168                                      |
| <b>test</b>          | 226                                   | 10                                  | 157                             | 174                                     | 94                                       |
| <b>true positive</b> | 210                                   | 2                                   | 130                             | 101                                     | 65                                       |
| <b>IoU</b>           | 0.80 (0.76-0.82)                      | 0.89 (0.88-0.90)                    | 0.89 (0.88-0.90)                | 0.80 (0.79-0.84)                        | 0.81 (0.76-0.82)                         |
| <b>mAP</b>           | 0.72                                  | 0.08                                | 0.83                            | 0.48                                    | 0.70                                     |

Table 2. IoU values for each lesion class.

IoU values represent the median for each class; values in parentheses indicate the 95% confidence intervals.

|            | <b>parenchyma</b> |             | <b>vessel</b> |             | <b>benign tumor</b> |             | <b>malignant tumor</b> |             |
|------------|-------------------|-------------|---------------|-------------|---------------------|-------------|------------------------|-------------|
| <b>IoU</b> | 0.62              | (0.48-0.67) | 0.9           | (0.80-0.91) | 0.81                | (0.75-0.86) | 0.83                   | (0.67-0.93) |

Table 3. Diagnostic performance of experts, trainees, and AI-assisted trainees.

Sensitivity, Specificity, Precision, and F1 Score represent the median values for each group; values in parentheses indicate the 95% confidence intervals.

|                    | <b>Sensitivity</b> |             | <b>Specificity</b> |             | <b>Precision</b> |             | <b>F1 Score</b> |             |
|--------------------|--------------------|-------------|--------------------|-------------|------------------|-------------|-----------------|-------------|
| <b>experts</b>     | 0.72               | (0.56-0.79) | 0.78               | (0.57-0.92) | 0.80             | (0.75-0.94) | 0.77            | (0.66-0.80) |
| <b>trainees</b>    | 0.55               | (0.25-0.68) | 0.22               | (0.12-0.79) | 0.53             | (0.42-0.75) | 0.53            | (0.32-0.63) |
| <b>trainees+AI</b> | 0.52               | (0.52-0.59) | 0.50               | (0.45-0.93) | 0.63             | (0.63-0.93) | 0.61            | (0.57-0.71) |

## Figures

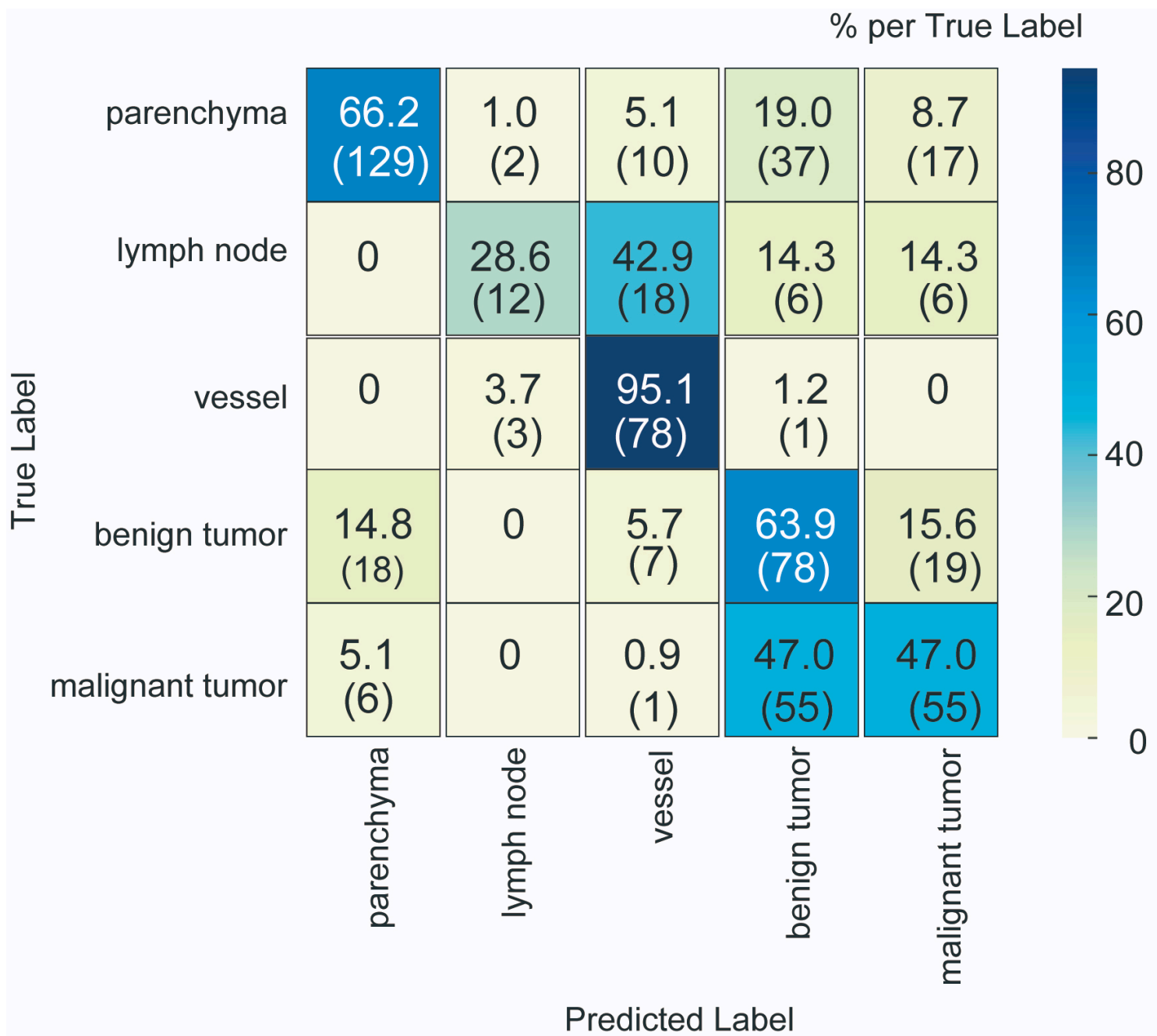


**Figure 1**

Row-normalized confusion matrix showing the percentage of predictions for each true class label. Each cell indicates the proportion (%) of samples with a given true label (row) that were predicted as the corresponding class (column). Numbers in parentheses represent the actual counts.

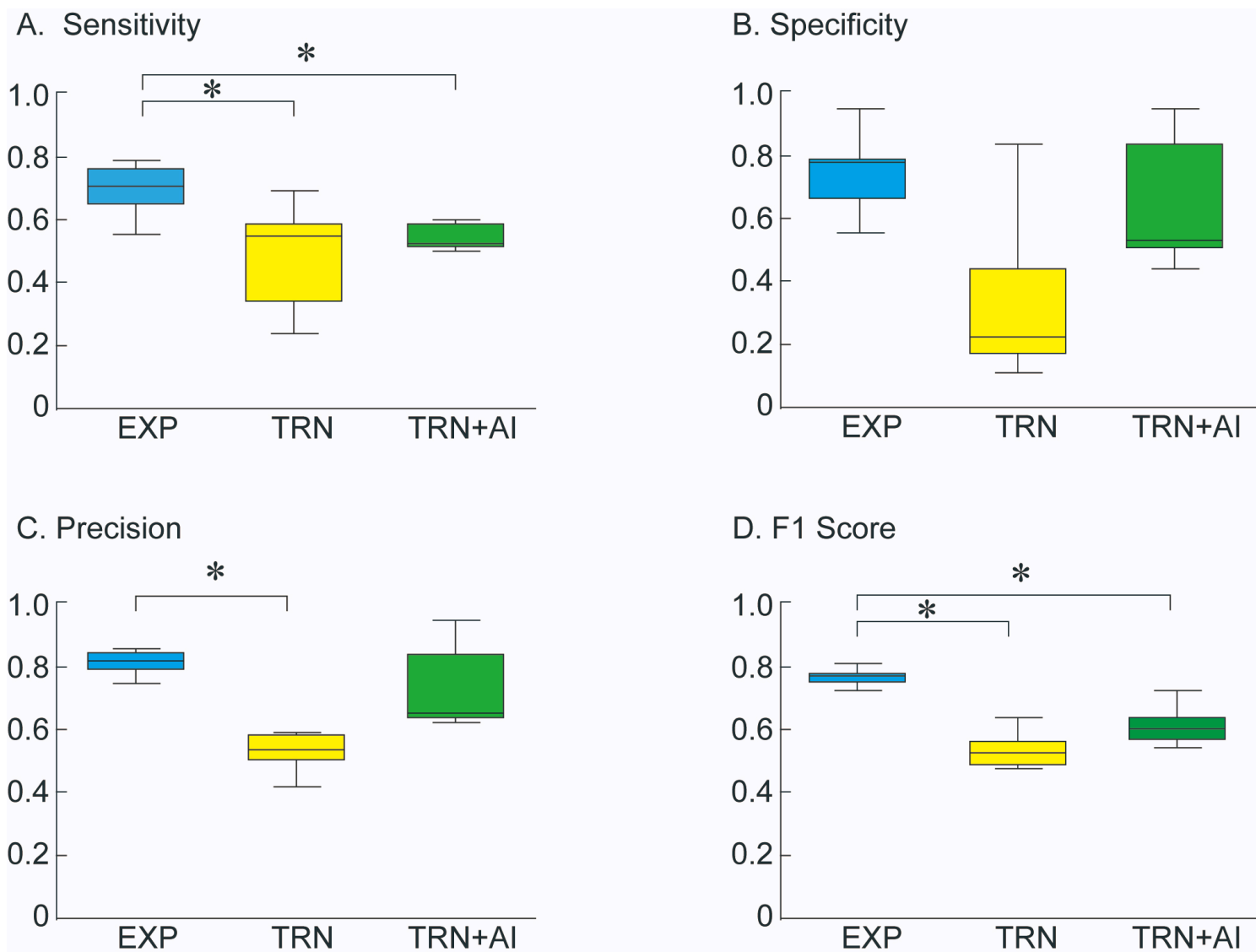


**Figure 2**

Boxplots with overlaid individual data points comparing diagnostic performance across three groups: expert doctors (EXP), trainees using AI assistance (TRN+AI), and students without AI assistance (TRN). Each panel shows one of the four performance metrics: A:Sensitivity, B:Specificity, C:Precision, and D:F1 Score. Boxes represent the interquartile range, with the median indicated by the central line. Whiskers extend to the range within  $1.5 \times \text{IQR}$ . Individual dots represent observed values per evaluator. \* indicate statistically significant differences at  $p < 0.05$