

# **Combining Image Similarity and Predictive AI Models to Decrease Subjectivity in Thyroid Nodule Diagnosis and Improve Malignancy Prediction**

Aishwarya Vedula

# Introduction

- Thyroid nodules are commonplace findings in clinical settings, with an estimated prevalence in the general population ranging from 4% to 6.5%.
- Currently, the best method to evaluate thyroid nodules involve ultrasound-guided fine-needle aspiration biopsy, which is invasive and can be emotionally distressing for patients.
- Up to 30% of biopsies lead to indeterminate results, requiring a repeat biopsy or surgery.
- AI has been increasingly utilized in various fields of medicine, including radiology and pathology, demonstrating its potential to augment the accuracy of diagnosis
- Image similarity assessment offers an efficient analysis of medical images that may exceed the capabilities of the human eye.
- The potential of combining AI-driven predictive models with image similarity assessment in thyroid nodule evaluation has not been explored for diagnosis and ACR TI-RADS assessment.

# Materials and Methods

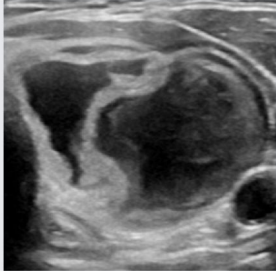
- In this study, we used software that integrates AI-driven predictive models with image similarity assessment for thyroid nodule evaluation.
- Two diverse datasets were used to evaluate the AI model.
  - The first dataset is an open-source dataset from Stanford University from 2021, which consists of 192 images of thyroid nodules collected between April 2017 and May 2018.
  - The second data set is from a private practice setting consisting of 118 thyroid nodule images from 2018 to 2023.
  - Both data sets had confirmed cytopathology and a TIRADS score.
- Inclusion criteria: Males and females, aged 18 years with thyroid surgery or biopsy at participating sites with a definitive diagnosis by cytology or pathology.
- Exclusion criteria: Patients below the age of 18 years; indeterminate thyroid nodules without a definitive diagnosis; ultrasound images of thyroid nodules containing annotations, markings, writings, or crosshair within the nodule and whole thyroid nodule not visible in the ultrasound section.

# Materials and Methods (cont.)

- The software uses static images in the AP dimension.
- It automatically identifies regions of interest.
- By comparing these regions to images in the training dataset, the software predicts whether the nodule is benign or malignant and also provides an ACR Thyroid Imaging Reporting and Data System (TI-RADS) score (Figure 1).
- Used Python language with Scikit-learn library to do the statistical analysis.

**SIMILAR IMAGES** 0.245

Region of interest



◀ 1/4 ▶

THE MOST SIMILAR IMAGE TO THE UPLOADED IMAGE IN OUR DATABASE HAS A DIAGNOSIS OF  
Benign

**TI-RADS DESCRIPTION**

TI-RADS	Prediction
Composition	Mixed cystic
Echogenicity	Isoechoic
Shape	Wider than tall
Margin	Ill defined
Echogenic Foci	No calcification

**TI-RADS SCORE AND RECOMMENDATION**

TR 2. Not suspicious, No biopsy indicated.

Figure 1

# Results

- In the Stanford public dataset, there were 17 malignant nodules and 175 benign nodules. The prevalence of malignancy in this dataset was 8 percent.
- In the private dataset there were 96 benign nodules and 22 malignant nodules. The prevalence of malignancy was 18.85%.

	Stanford Data	Private Data
Sensitivity	1	0.91
Specificity	0.55	0.95
PPV	0.18	0.8
NPV	1	0.98
AUC ROC	0.78	0.93

# Results (cont.)

- The Pearson correlation coefficient between ground truth cytopathology diagnosis and AI diagnosis was 0.824 with a p-value of  $2.29 \times 10^{-31}$ , indicating a strong positive correlation that is statistically significant.
- The AI program and ground truth diagnoses exhibit high agreeability with a concordance rate of 94.26 percent and an F1 score of 85.21 percent.
- Regarding the TI-RADS score by a physician vs that was predicted by AI algorithm, the Pearson Correlation Coefficient was 0.877 with  $p < 0.001$  indicating a strong linear relationship between the two readings.
- Cohen's Kappa for physician readings vs AI reading was 0.753. This indicates substantial agreement between the physician and the AI system.

# Discussion

- AI can be particularly useful in evaluating thyroid nodules, typically for risk stratification.
- Recent studies suggested that the performance of artificial intelligence models was better or at par with radiologists.
  - These studies postulated that artificial intelligence software can be especially beneficial for physicians with less experience.
- Currently, the United States Food and Drug Administration has approved four AI platforms for thyroid disease.
- Despite the reported efficacy of artificial intelligence, common concerns exist with its usability.
- Our software addresses some of these concerns. By generating human-understandable descriptors and explanations for its decisions, our software's interpretations can be verified by physicians.

# Discussion (cont.)

- There is a need for explainability in medical AI models.
- Explainable AI is a set of tools and methods that help people understand and interpret predictions made by their machine learning algorithms.
- This consists of an explainable model and an explanation interface so human users can understand what caused the model to make a certain conclusion or prediction.
- Focusing on the end user as well as interventions to increase the understanding of a medical algorithmic decision process would be important to increase utilization.
  - This is especially crucial in medicine because medical professionals need to understand the basis for an algorithm's diagnosis.
  - Our AI algorithm took these factors into consideration with its easy to use interface and transparency in decision making



# Discussion (cont.)

- External validation is necessary to avoid the consequences of a model with low adaptability.
- It is critical to understanding the clinical utility of prediction models.
- Hence we undertook external validation on two widely different datasets and demonstrated good performance.
- One of the unique aspects of our research is its integration of image similarity assessment and TI-RADS scoring to produce diagnostic outcomes, a combination that has not been explored before in thyroid nodules.
- Our software identifies similar images from its database when compared to the test image. The diagnosis of the most similar image is displayed as the output of the AIbX algorithm. A TI-RADS score description and recommendation is then produced by the model to enable verification by medical professionals.

# Discussion (cont.)

- Some limitations of our study were the small sample size, use of static images, and the low number of malignant cases.
- These could have contributed to the low positive predictive value.
- In the future, we could test it on databases with a higher prevalence of malignancy, but the average prevalence of malignancy in the combined dataset was very similar to the general population. Furthermore, this software was not prospectively evaluated in a clinical setting.
- The results from the study showed a high negative predictive value, meaning if our algorithm predicted that a nodule is benign, it had a very low probability of being malignant.
- The AI algorithm missed only 2 malignant nodules.
- It performed well with ultrasound images across multiple institutions using different ultrasound machines and showed no bias across nodules of various types and sizes and age groups.

# Conclusion

- The combined image similarity and AI model demonstrated high negative predictive value with a potential for a 60% reduction in the need for biopsy.
- This holds significant clinical implications, as the integration of image similarity and AI-driven predictive models could revolutionize the approach to thyroid nodule evaluation.
- This paves the way for non-invasive screening and also has the potential to greatly reduce the burden on patients and healthcare costs.

# References

Please view paper submission.