Abstract

Objectives

To evaluate the efficacy of combining predictive artificial intelligence and image similarity model to risk stratify thyroid nodules, using a retrospective study.

Methods

Two datasets were used to determine the efficacy of the algorithm. One was the publicly available Stanford dataset consisting of ultrasound images of 192 nodules between April 2017 to May 2018 and the second one was from a private practice setting consisting of 118 thyroid nodule images from 2018-2023. All the nodules had definitive diagnosis either by biopsy or by surgery. The software was used to predict the diagnosis and TI-RADS score.

Results

In the Stanford dataset, the AI algorithm predicted malignancies with a sensitivity of 1.0 and a specificity of 0.55. The PPV was 0.18 and the NPV was 1.0. The AUCROC was 0.78. The AI algorithm did not miss any cases of cancer. TI-RADS based clinical recommendation had a polychoric correlation of 0.67. In the private dataset, the AI algorithm predicted malignancies with a sensitivity of 0.91 and a specificity of 0.95. The PPV was 0.8 and NPV was 0.98. AUCROC was 0.93 and accuracy was 0.94. TI-RADS based clinical recommendation had a polychoric correlation of 0.94 for this dataset.

Conclusion

The AI model demonstrated high negative predictive value with a potential for 60% reduction in the need for biopsy. This could reduce the burden on patients and healthcare costs.

Introduction

Thyroid nodules are commonplace findings in clinical settings, with an estimated prevalence in the general population ranging from 4% to 6.5%.[1] Though the vast majority of these nodules are benign, roughly 10%-15% of them are malignant.[2] 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.[3] Moreover, up to 30% of biopsies lead to indeterminate results, requiring a repeat biopsy or surgery.[4] To effectively distinguish whether a thyroid nodule is benign or malignant is crucial in determining accurate clinical management and reducing the number of unnecessary biopsies.

AI has been increasingly utilized in various fields of medicine, including radiology and pathology, demonstrating its potential to augment the accuracy of diagnosis.[5] In thyroid nodule evaluation particularly, AI-driven predictive models offer non-invasive strategies to detect malignancies.[6]

Additionally, image similarity assessment, which involves the comparison of visual characteristics of images, can also be used in medical diagnostics. It offers an efficient analysis of medical images that may exceed the capabilities of the human eye.[7] 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.[8] Therefore, in this paper we are evaluating the benefits of combining these methods. We elucidate the methods of our software, evaluate its performance, and discuss the potential implications of combining image similarity and AI to provide better screening for thyroid nodules.

Materials and methods

In this study, we used software that integrates AI-driven predictive models with image similarity assessment for thyroid nodule evaluation. Version 2 of this software also predicts ACR TI-RADS. PEARL IRB determined the study to be exempt. 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 .[9] These images were 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. This data set consists of images from an in-house thyroid ultrasound machine as well as an external radiology ultrasound machine. Both data sets had confirmed cytopathology and a TIRADS score. For the second data set with in-house images, the TIRADS score was assigned by the performing endocrinologist (RV) which was then reviewed and confirmed by a second endocrinologist (RP). Any discrepancies were resolved by a third endocrinologist(JC).

The inclusion criteria for the study were males and females, aged 18 years with thyroid surgery or biopsy at participating sites with a definitive diagnosis by cytology or pathology. Indeterminate nodules (Bethesda III, IV, and V) upon initial evaluation should have undergone surgery with a definitive diagnosis to be included in the study. Thyroid nodules measuring between 5 mm and 40 mm (4.0cm) in the maximum dimension by ultrasound imaging in transverse dimension. The longest diameter of the thyroid nodule should be less than the length of the ultrasound transducer.

Exclusion criteria for the study were 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; metastasis to the thyroid from other malignancies as well as lymphoma of the thyroid were also excluded; multinodular goiters without a clearly separable nodule on ultrasound images and nodules that underwent radioactive iodine treatment, ethanol ablation, radiofrequency ablation or laser ablation.

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).

We used Python language with Sckit-learn library to do the statistical analysis.[10]

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. Compared to ground truth, the AI algorithm predicted malignancies with a sensitivity of 1.0 and a specificity of 0.55. The positive predictive value (PPV) was 0.18 and the negative predictive value (NPV) was 1.0. The AUCROC was 0.78. The AI algorithm did not miss any cases of cancer. ACR TI-RADS based clinical recommendation had a polychoric correlation of 0.67.

In the private dataset there were 96 benign nodules and 22 malignant nodules. The prevalence of malignancy was 18.85%. In this dataset, the AI algorithm predicted malignancies with a sensitivity of 0.91 and a specificity of 0.95. The PPV was 0.8 and NPV was 0.98. AUCROC was 0.93 and accuracy was 0.94. In this dataset, 5 out of the 99 benign nodules were read as malignant by the AIBx algorithm. These nodules had high risk features with TI-RADS scores 4-5 for 4 out of the 5 nodules and TI-RADS score of 3 for 1 out of the 5 nodules. The AIBx algorithm also predicted 2 out of the 22 malignant nodules as benign. These nodules had a TI-RADS score of 3. TI-RADS-based clinical recommendation had a polychoric correlation of 0.94 for this dataset. Table 1, shows comparison of AI predictions on both datasets.

The Pearson correlation coefficient between ground truth cytopathology diagnosis and AI diagnosis was 0.824 with a p-value of 2.29 x 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

In recent years, artificial intelligence tools have become increasingly prevalent across multiple disciplines.

AI can be particularly useful in evaluating thyroid nodules, typically for risk stratification.[6,11]

Recent studies suggested that the performance of artificial intelligence models was better or at par with radiologists.·[12,13] 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, such as the proper integration of AI and radiologist interpretations and assessment of productivity. Furthermore, the authors concluded that the successful adoption of AI platforms requires that the software be incorporated into the physician's workflow seamlessly and should have external validation studies.[6] 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. Having a high negative predictive value and decreasing biopsy need by 60%, this software demonstrated its ability to reduce healthcare spending. Coupled with its easy-to-use nature, this software ensures practicality, workflow efficiency, and demonstrable performance, all of which are critical for acceptance in clinical settings.

Need for explainability in medical AI models.

Explainable AI or interpretable AI, is a set of tools and methods that help people understand and interpret predictions made by their machine learning algorithms.[14] 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, which helps characterize model accuracy, fairness, transparency, and outcomes in decision-making powered by AI.[15] However, there is a reluctance to use medical AI due to a combination of lack of focus on the end-user by developers of the AI leading to a subjective difficulty of understanding the algorithm and more comfort with human decision making. [16,17] Therefore, focusing on the end user by developers of medical AI 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. A false negative could mean that a patient doesn't receive life-saving treatment, and a false positive could result in a patient receiving expensive and invasive treatments when it isn't necessary to do so.[18] A level of explainability is essential for medical professionals to have comfort in integrating medical AI into practice. Our AI algorithm took these factors into consideration with its easy to use interface and transparency in decision making that makes it very user-friendly and easy to integrate into daily practice with confidence.

Due to a lack of validation, many AI technologies are not applied in clinical decision-making 15. External validation is used to evaluate predictive capabilities for target clinical implementations in different populations and settings.[19] Predictive models often perform well under training datasets. However, there is a discrepancy between training and validation performance. This discrepancy even appears when training and validation datasets are from the same populations and settings. Poorly developed models lead to exacerbated disparities in healthcare provisions and outcomes. Thus, external validation is necessary to avoid the consequences of a model with low adaptability. External validation is critical to understanding the clinical utility of prediction models.[20] 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. Image similarity assessment uses visual pattern recognition to compare and contrast features of a nodule against a repository of images already classified as malignant or benign. This results in a more accurate evaluation while simultaneously allowing medical professionals to verify the algorithm's conclusions. A TI-RADS score aids in this endeavor by providing human-understandable descriptors to fill the gap between the novelty of AI algorithms and the traditional use of clinical assessment. 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.

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. This would translate into observation as opposed to undergoing a biopsy. The AI algorithm missed only 2 malignant nodules. Both of these nodules were follicular carcinomas of the thyroid and had benign characteristics isoechoic, clear borders, and small central cystic spaces. However, feedback to the AI with these types of nodules as malignant could lead to better predictions in the future. Our AI model 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. Not only does this pave the way for non-invasive screening, but it also has the potential to greatly reduce the burden on patients and healthcare costs alike.

References

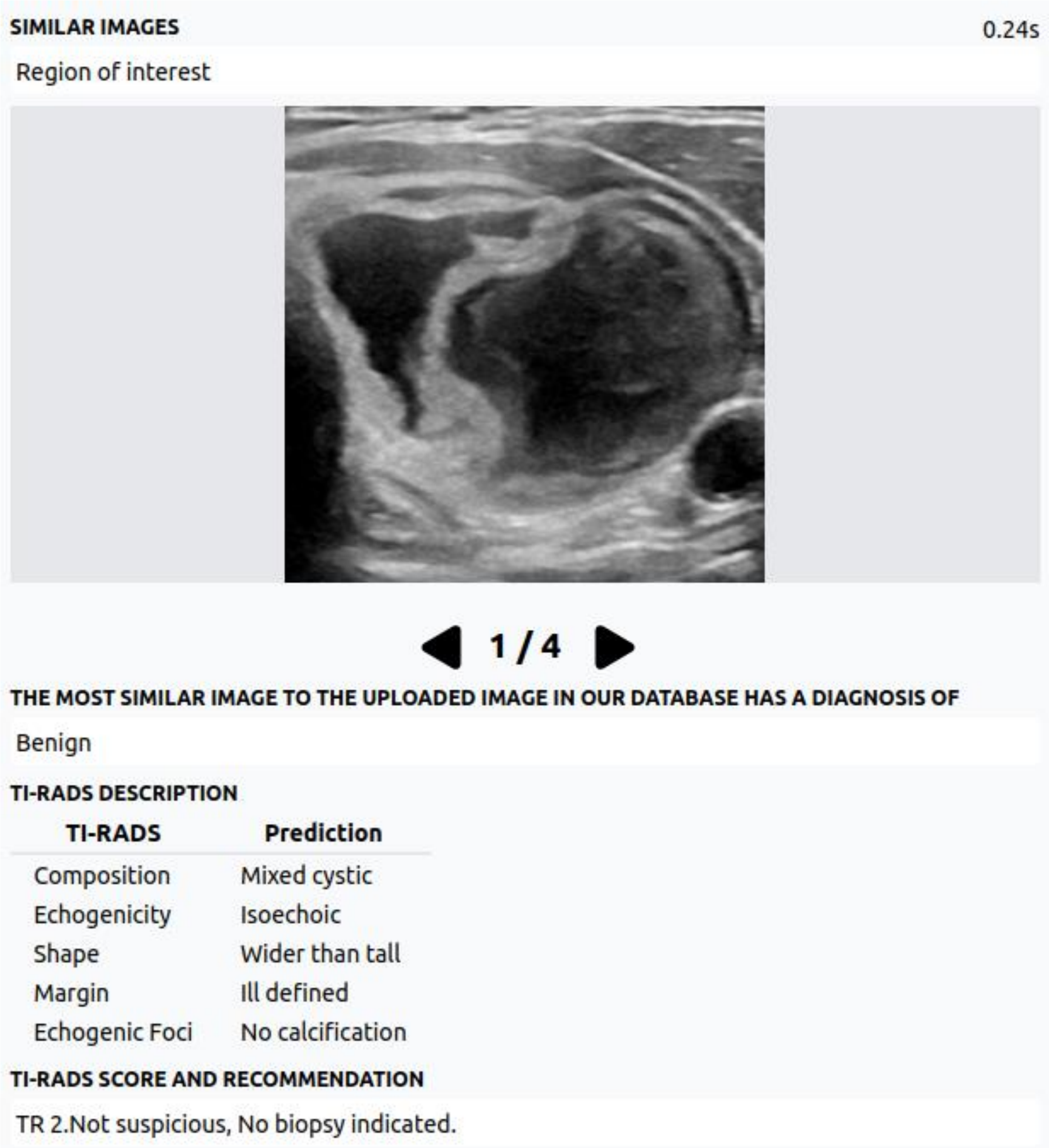
1. Popoveniuc G, Jonklaas J. Thyroid Nodules. *Med Clin North Am*. 2012;96(2):329-349. doi:10.1016/j.mcna.2012.02.002
2. Kamran SC, Marqusee E, Kim MI, et al. Thyroid nodule size and prediction of cancer. *J Clin Endocrinol Metab*. 2013;98(2):564-570. doi:10.1210/jc.2012-2968
3. Jasim S, Dean DS, Gharib H. Fine-Needle Aspiration of the Thyroid Gland. In: Feingold KR, Anawalt B, Blackman MR, et al., eds. *Endotext*. MDText.com, Inc.; 2000. Accessed February 19, 2024. http://www.ncbi.nlm.nih.gov/books/NBK285544/
4. Yip L, Farris C, Kabaker AS, et al. Cost Impact of Molecular Testing for Indeterminate Thyroid Nodule Fine-Needle Aspiration Biopsies. *J Clin Endocrinol Metab*. 2012;97(6):1905-1912. doi:10.1210/jc.2011-3048
5. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7
6. Tessler FN, Thomas J. Artificial Intelligence for Evaluation of Thyroid Nodules: A Primer. *Thyroid Off J Am Thyroid Assoc*. 2023;33(2):150-158. doi:10.1089/thy.2022.0560
7. Krupinski EA. Current perspectives in medical image perception. *Atten Percept Psychophys*. 2010;72(5):10.3758/APP.72.5.1205. doi:10.3758/APP.72.5.1205
8. Tessler FN, Middleton WD, Grant EG, et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. *J Am Coll Radiol JACR*. 2017;14(5):587-595. doi:10.1016/j.jacr.2017.01.046
9. Yamashita R, Kapoor T, Alam MN, et al. Toward Reduction in False-Positive Thyroid Nodule Biopsies with a Deep Learning-based Risk Stratification System Using US Cine-Clip Images. *Radiol Artif Intell*. 2022;4(3):e210174. doi:10.1148/ryai.210174
10. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(85):2825-2830.
11. Wildman-Tobriner B, Taghi-Zadeh E, Mazurowski MA. Artificial Intelligence (AI) Tools for Thyroid Nodules on Ultrasound, From the AJR Special Series on AI Applications. *AJR Am J Roentgenol*. 2022;219(4):1-8. doi:10.2214/AJR.22.27430

12. Park VY, Han K, Seong YK, et al. Diagnosis of Thyroid Nodules: Performance of a Deep Learning Convolutional Neural Network Model vs. Radiologists. *Sci Rep*. 2019;9(1):17843. doi:10.1038/s41598-019-54434-1
13. He LT, Chen FJ, Zhou DZ, et al. A Comparison of the Performances of Artificial Intelligence System and Radiologists in the Ultrasound Diagnosis of Thyroid Nodules. *Curr Med Imaging*. 2022;18(13):1369-1377. doi:10.2174/1573405618666220422132251
14. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115. doi:10.1016/j.inffus.2019.12.012
15. What is Explainable AI? - Unite.AI. Accessed February 19, 2024. https://www.unite.ai/what-is-explainable-ai/
16. Cadario R, Longoni C, Morewedge CK. Understanding, explaining, and utilizing medical artificial intelligence. *Nat Hum Behav*. 2021;5(12):1636-1642. doi:10.1038/s41562-021-01146-0
17. Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *Npj Digit Med*. 2022;5(1):1-15. doi:10.1038/s41746-022-00699-2
18. McNamara M. Explainable AI: What is it? How does it work? And what role does data play? Published February 22, 2022. Accessed February 19, 2024. https://www.netapp.com/blog/explainable-ai/
19. Tsopra R, Fernandez X, Luchinat C, et al. A framework for validating AI in precision medicine: considerations from the European ITFoC consortium. *BMC Med Inform Decis Mak*. 2021;21(1):274. doi:10.1186/s12911-021-01634-3
20. Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ*. 2024;384:e074820. doi:10.1136/bmj-2023-074820

Table 1: Comparison of AI predictions on both datasets

	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

Figure 1: AI software result interface.



SIMILAR IMAGES 0.24s

Region of interest

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.