


Thy-Wise: An interpretable machine learning model for the evaluation of thyroid nodules

Zhe Jin¹ | Shufang Pei² | Lizhu Ouyang³ | Lu Zhang¹ | Xiaokai Mo¹ | Qiuying Chen¹ | Jingjing You¹ | Luyan Chen¹ | Bin Zhang¹ | Shuixing Zhang¹ 

¹Department of Radiology, The First Affiliated Hospital of Jinan University, Guangzhou, Guangdong, China

²Department of Ultrasound, Guangdong Provincial People's Hospital/Guangdong Academy of Medical Sciences, Guangzhou, Guangdong, China

³Department of Ultrasound, Shunde Hospital of Southern Medical University, Foshan, China

Correspondence

Bin Zhang and Shuixing Zhang, The First Affiliated Hospital of Jinan University, No. 613 Huangpu West Road, Tianhe District, Guangzhou, Guangdong 510627, China. Email: binzhang1988@jnu.edu.cn (B. Z.) and zxs7515@jnu.edu.cn (S. Z.)

Funding information

China Postdoctoral Science Foundation, Grant/Award Number: 2016M600145; Guangdong Grand Science and Technology Special Project, Grant/Award Number: 2015B010106008; National Scientific Foundation of China, Grant/Award Number: 81571664; Science and Technology Planning Project of Guangdong Province, Grant/Award Numbers: 2014A020212244, 2016A020216020; the Scientific Research General Project of Guangzhou Science Technology and Innovation Commission, Grant/Award Number: 201605110912158

Abstract

Current risk stratification systems for thyroid nodules suffer from low specificity and high biopsy rates. Recently, machine learning (ML) is introduced to assist thyroid nodule diagnosis but lacks interpretability. Here, we developed and validated ML models on 3965 thyroid nodules, as compared to the American College of Radiology Thyroid Imaging, Reporting and Data System (ACR TI-RADS). Subsequently, a SHapley Additive exPlanation (SHAP) algorithm was leveraged to interpret the results of the best-performing ML model. Clinical characteristics including thyroid-function tests were collected from medical records. Five ACR TI-RADS ultrasonography (US) categories plus nodule size were assessed by experienced radiologists. Random forest (RF), support vector machine (SVM) and extreme gradient boosting (XGBoost) were used to build US-only and US-clinical ML models. The ML models and ACR TI-RADS were compared in terms of diagnostic performance and unnecessary biopsy rate. Among the ML models, the US-only RF model (hereafter, Thy-Wise) achieved the optimal performance. Compared to ACR TI-RADS, Thy-Wise showed higher accuracy (82.4% vs 74.8% for the internal validation; 82.1% vs 73.4% for external validation) and specificity (78.7% vs 68.3% for internal validation; 78.5% vs 66.9% for external validation) while maintaining sensitivity (91.7% vs 91.2% for internal validation; 91.9% vs 91.1% for external validation), as well as reduced unnecessary biopsies (15.3% vs 32.3% for internal validation; 15.7% vs 47.3% for external validation). The SHAP-based interpretation of Thy-Wise enables clinicians to better understand the reasoning behind the diagnosis, which may facilitate the clinical translation of this model.

KEYWORDS

diagnosis, machine learning, random forest, thyroid nodules, ultrasonography

What's new?

Low specificity of existing methods of thyroid nodule assessment for thyroid cancer can lead to unnecessary biopsy and overtreatment. Here, the authors constructed an interpretable machine learning model, called Thy-Wise, based on a SHapley Additive exPlanation algorithm to evaluate thyroid nodules. Thy-Wise achieved greater accuracy and specificity at similar sensitivity when

Abbreviations: ACR, American College of Radiology; AUC, area under the curve; FNAB, fine-needle aspiration biopsy; ML, machine learning; NPV, negative predictive value; PPV, positive predictive value; RF, random forest; ROC, receiver operating characteristic; SHAP, SHapley Additive exPlanation; SVM, support vector machine; TI-RADS, Thyroid Imaging Reporting and Data System; TR, TI-RADS risk; US, ultrasonography; XGBoost, extreme gradient boosting.

Zhe Jin, Shufang Pei and Lizhu Ouyang contributed equally to this work.

compared to the American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS). Thy-Wise also had a decreased unnecessary biopsy rate. The application of methods to explain model output may help improve diagnostic interpretations of thyroid nodules, rendering Thy-Wise a valuable tool for clinical decision making.

1 | INTRODUCTION

Thyroid nodules are extremely common and can be detected in more than 60% of asymptomatic adults in the general population.¹ To date, ultrasonography (US) remains the preferred diagnostic imaging to screen thyroid nodules.² The widespread application of US has markedly increased the detection of thyroid nodules, which essentially facilitates early intervention of thyroid cancer.³ Several early risk-stratification systems have been designed for clinical management of thyroid nodules; however, they depend on operator experience or local practice patterns.⁴ In 2017, the American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS) was released to help radiologists improve consistency and enhance report quality.⁵ In the ACR TI-RADS, a nodule's TI-RADS risk (TR) level was determined by the total point summed up from all five gray-scale US categories, which underlies the decision of no further evaluation, fine-needle aspiration biopsy (FNAB) or surveillance in addition to nodule size.⁶

The application of ACR TI-RADS contributes to a decrease in the FNAB rate and better management recommendations for thyroid nodules.⁷ Previous studies suggested that ACR TI-RADS had a much lower unnecessary FNAB rate than other systems, such as American Thyroid Association (ATA) guidelines for the assessment of thyroid nodules, European Thyroid Association TI-RADS and Korean Society of TI-RADS.⁸⁻¹³ However, given the high prevalence of thyroid nodules and the low incidence of thyroid cancer, the low specificity of ACR TI-RADS (44.0%-67.3%) may result in a large proportion of patients without clinically significant nodules undergoing unnecessary biopsies or even diagnostic lobectomy.^{7,14,15} A high rate of malignancies does not receive a recommendation for FNAB using ACR TI-RADS (17%-32%) compared to other guidelines.¹⁶ ACR TI-RADS stratifies nodules into five risk categories, whereas the range of category 5 (TR5) is so wide (20%-100%) that against individualized strategies.⁵ Thus, a noninvasive and low-cost approach with a higher specificity is desirably needed to determine the requirement of FNAB.

Machine learning (ML) approaches use statistical methods that can identify the underlying pattern and classification from medical data, and have been successfully applied to improve patient care.¹⁷ Recently, ML techniques have been introduced to screen thyroid cancer.¹⁸ By characterizing the morphological and textural features on the US images, ML showed similar or even better performance than experienced radiologists.¹⁹⁻²² Nevertheless, such ML models may require explanations because they are applied in a high-risk clinical scenario, meaning a mistake may have adverse consequences (eg, biopsy or surgery). Although published ML models showed enough classification accuracy, their utilization in clinical practice are

challenging because their outputs are hard to be interpreted, like a "black box," and hence not actionable.²³ A human-interpretable model is crucial for maintaining the likelihood of data-driven knowledge discovery, which explains why a specific prediction was given.²⁴ This lack of explainability has thus far hindered the application of powerful methods including deep learning (DL, a subfield of ML), radiomics and ensemble models in clinical decision support of thyroid nodules.

To overcome these shortcomings, we aimed to develop the best performing and interpretable ML model to evaluate the malignancy of thyroid nodules in a large-scale population. We hypothesized our ML model can achieve higher specificity than ACR TI-RADS at similar sensitivity and reduce unnecessary FNAB. Additionally, we also verified the added value of clinical data including thyroid function to the ML model.

2 | MATERIALS AND METHODS

2.1 | Study population

Between January 2013 and December 2018, a total of 3036 consecutive patients aged 18 years old or older with thyroid nodules (≥ 10 mm in maximum diameter) who underwent FNAB and/or thyroidectomy in two tertiary hospitals, were screened. The inclusion criteria were as follows: (1) patients had definitive benign or malignant thyroid nodules with confirmed histology or cytology (Bethesda category); (2) patients had available clinical data, including demographics, underlying thyroid disease and biochemical findings of thyroid function; and (3) gray-scale US was performed within 1 month prior to FNAB or surgery. Patients were excluded due to inconclusive diagnoses of pathological findings, incomplete clinical data or US images. Finally, 3098 nodules from 2848 patients were included as the primary dataset. The clinical characteristics were collected from medical records, including age, sex, underlying thyroid diseases (ie, Hashimoto's thyroiditis, Grave's disease or none), thyrotropin (TSH) (normal range, 0.34-5.6 μ U/mL), free triiodothyronine (T3) (normal range, 3.8-6 pmol/L), free thyroxine (T4) (normal range, 7.5-21.1 pmol/L), total triiodothyronine (TT3) (normal range, 1.34-2.73 nmol/L) and total thyroxine (TT4) (normal range, 78.4-157.4 nmol/L) at admission.

Between May 2016 and November 2019, an independent validation dataset of 867 nodules in 765 consecutive patients aged 18 years old or older with thyroid nodules (≥ 10 mm in maximum diameter), was identified using the same inclusion and exclusion criteria as the primary dataset. Figure 1 illustrates the flowchart of inclusion criteria for initial population and exclusion criteria for the final study population.

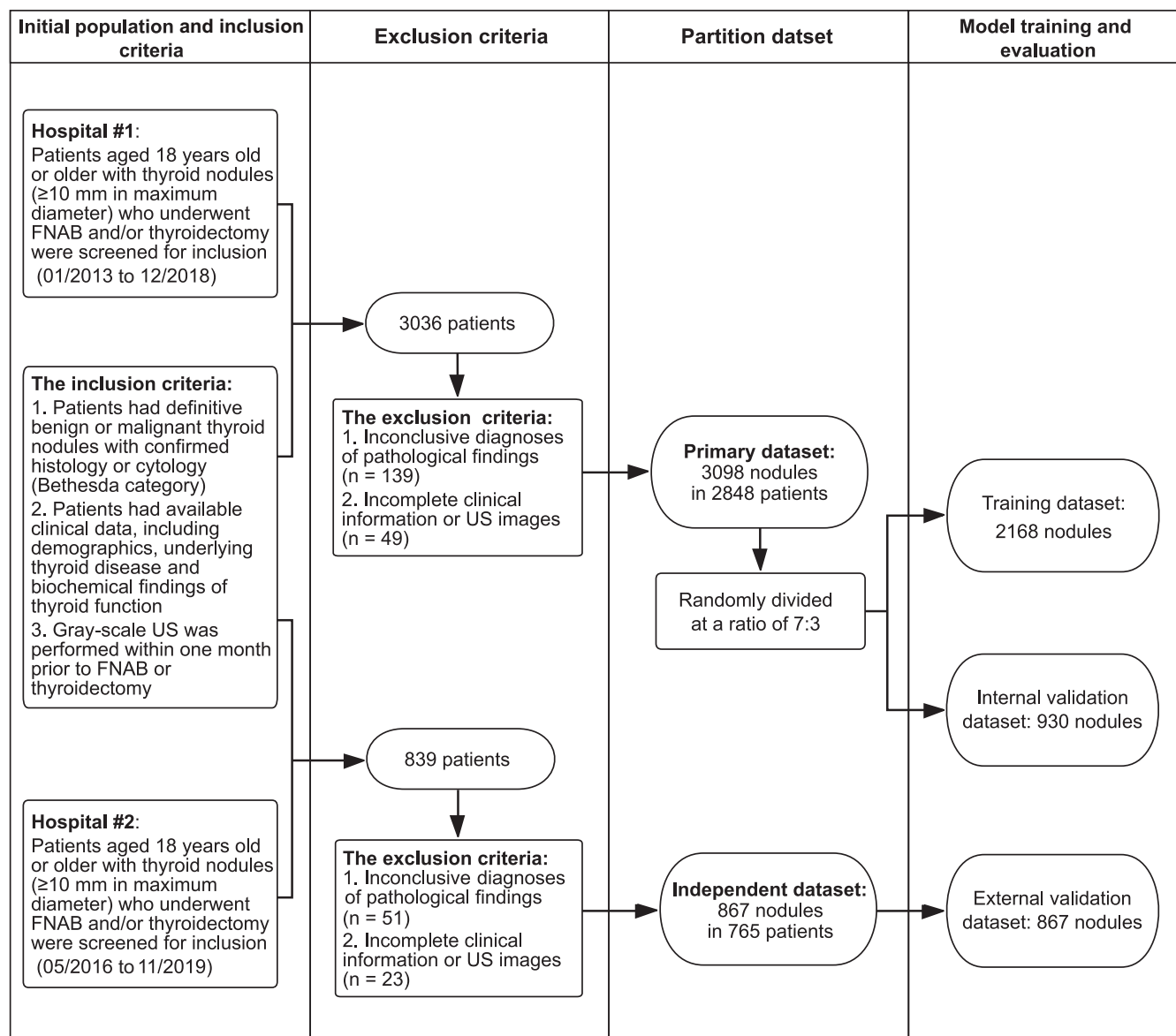


FIGURE 1 A flowchart of patient and nodule inclusion and exclusion. A total of 3965 thyroid nodules from two centers were finally reviewed. The primary dataset from Hospital #1 was randomly divided into a training dataset (n = 2168) and an internal validation dataset (n = 930) by a ratio of 7:3, and the independent dataset from Hospital #2 was served as an external validation dataset (n = 867)

2.2 | Ultrasonography examination

Thyroid US scans were performed with 6–13 MHz linear probes of real-time US systems. Patients from the primary dataset were examined with HI Vision Ascendus (Hitachi, Tokyo, Japan), HI Vision 900 (Hitachi, Tokyo, Japan), TUS-A500 (Toshiba, Tokyo, Japan) and Resona 8 (Mindray Bio-Medical Electronic Co., Shenzhen, China). Patients from the external validation dataset were examined with HI Vision Preirus (Hitachi, Tokyo, Japan), HI Vision 900 (Hitachi, Tokyo, Japan), TUS-A400 (Toshiba, Tokyo, Japan) and Resona 8 (Mindray Bio-Medical Electronic Co., Shenzhen, China). The acquisition of US images adhered to American College of Radiology accreditation standards, and these images were stored on the picture

archiving and communication system (GE Centricity; GE Healthcare, Milwaukee, WI).

2.3 | Ultrasonography image analysis

US images were evaluated in a blinded fashion, with no US or pathology reports available, by two board-certified radiologists (with more than 10 years of experience in thyroid sonography) independently. Before our study started, they reached a consensus on a feature-by-feature basis as the “truth” US features of a thyroid nodule. All the annotations in the images and clips were eliminated before review.

Nodule size was measured as the maximal dimension on US images and the five gray-scale US categories were reviewed according to the ACR TI-RADS lexicon⁵: composition, echogenicity, shape, margin and echogenic foci. In the ACR TI-RADS, the TR level for nodules was determined by the total score of the five US categories, ranging from TR1 (benign) to TR5 (highly suspicious).

2.4 | Machine learning model development and evaluation

The nodules from the primary dataset were randomly divided into a training dataset and an internal validation dataset at a ratio of approximately 7:3. We developed both US-only and US-clinical ML models. The US-only ML models contained six US features: composition, echogenicity, shape, margin, echogenic foci and nodule size. The US-clinical ML models combined the six US features with significant clinical variables. Univariate and multivariate analyses were performed to identify the independent clinical risk factors associated with malignancy. Three different ML algorithms namely, random forest (RF), support vector machine (SVM) and extreme gradient boosting (XGBoost)

were utilized to develop diagnostic models. Briefly, RF was an ensemble algorithm that integrates the output of many decision trees to reach a single output.²⁵ XGBoost was an optimized distributed gradient boosting library that converts a group of weak learners into strong learners.²⁵ SVM was an algorithm that identifies high-dimensional bounds and classifies data points explicitly by determining the maximum-margin hyperplane.²⁵ The ML models were built via 10-fold cross-validation on the training dataset, whose optimal hyperparameters are shown in Table S1. The ML models were then internally validated and externally tested. The code used to build ML models is available at <https://github.com/MedAI-Lab/Thy-Wise>.

The diagnostic performance to assess thyroid nodules and unnecessary biopsy rate were compared between the ML models and ACR TI-RADS. Their diagnostic performance was evaluated by the receiver operating characteristic (ROC) curve analysis. The area under the curve (AUC) was used to quantitatively measure discrimination. Corresponding two-by-two confusion matrices with the number of true positive, false positive, false negative and true negative values were also plotted. The unnecessary biopsy rate was defined as the percentage of FNAB-indicated benign nodules to the total number of nodules. The biopsy rate of malignancy was defined as the percentage of

TABLE 1 Clinical and histopathologic characteristics of patients and nodules

Characteristics	Training dataset	Internal validation dataset	External validation dataset
No. of patients	2041	807	765
Sex			
Female	1325 (64.9)	526 (65.2)	578 (75.6)
Male	716 (35.1)	281 (34.8)	187 (24.4)
Age (years), mean \pm SD	46.1 \pm 12.6	46.5 \pm 13.2	47.5 \pm 13.0
Underlying thyroid diseases			
Hashimoto's thyroiditis	152 (7.4)	61 (7.6)	33 (4.3)
Grave's disease	64 (3.1)	16 (2.0)	27 (3.5)
None	1825 (89.4)	730 (90.5)	705 (92.2)
T3 (ng/mL)	4.8 \pm 0.9	4.7 \pm 0.8	5.1 \pm 1.1
T4 (μ g/dL)	12.7 \pm 5.0	12.4 \pm 4.4	12.4 \pm 4.2
TT3 (ng/dL)	1.9 \pm 0.5	1.9 \pm 0.5	1.6 \pm 0.6
TT4 (μ g/dL)	111.1 \pm 24.2	110.2 \pm 24.3	109.0 \pm 28.7
TSH (μ IU/mL)	2.8 \pm 1.7	2.9 \pm 1.8	1.6 \pm 2.7
No. of nodules	2168	930	867
Nodule size (mm), mean \pm SD	27.7 \pm 8.7	27.5 \pm 8.6	29.3 \pm 11.7
No. of malignant nodules	586 (27.1)	264 (28.3)	235 (27.2)
Papillary carcinomas	563 (96.1)	248 (93.9)	226 (96.2)
Follicular carcinomas	21 (3.6)	15 (5.7)	8 (3.4)
Medullary carcinomas	2 (0.3)	1 (0.4)	1 (0.4)
No. of benign nodules	1582 (72.9)	666 (71.6)	632 (72.8)
Nodular hyperplasia	1404 (88.7)	592 (88.9)	567 (89.7)
Follicular adenoma	118 (7.5)	51 (7.7)	41 (6.5)
Thyroiditis	60 (3.8)	23 (3.5)	24 (3.8)

Note: Data are expressed as number of patients (percentage) or mean \pm SD.

Abbreviations: T3, free triiodothyronine; T4, free thyroxine; TSH, thyrotropin; TT3, total triiodothyronine; TT4, total thyroxine.

FNAB-indicated malignant nodules to the total number of malignant nodes. Two aforementioned radiologists determined by consensus whether a nodule required a biopsy according to ACR TI-RADS recommendation. FNAB was indicated when one or more nodules were⁵: (1) classified as TR3 with a nodule size ≥ 2.5 cm; (2) classified as TR4 with a nodule size ≥ 1.5 cm; and (3) classified as TR5 with a nodule size ≥ 1.0 cm. To maintain consistency with the FNAB recommendation by ACR TI-RADS, the nodules were not indicated for FNAB if the output of the ML model was negative (below the cutoff value), whereas they were indicated for FNAB if the output was positive (above the cutoff value).²⁶

2.5 | Interpretability of the best-performing machine learning model

The local and global interpretability of black-box ML predictions is meaningful but challenging. One of the key factors that determine whether ML models can be accepted by clinicians is if they can comprehend how the ML models “think.” The higher the transparency of the ML models, the easier it is for clinicians to accept and thus make

an appropriate clinical decision that benefits patient care.²⁷ In our study, we explained the optimal ML model using the SHAP algorithm.²⁸ The SHAP explanation algorithm inspired by coalitional game theory-individual feature value of a data instance act like players in a coalition (diagnosis task) and the Shapley values let us know how to fairly distribute the “payout” (diagnostic performance) among the features.²⁹ The goal of SHAP is to explain the diagnosis by computing the contribution of each feature value to the prediction.²⁹ The SHAP method is preferred over the other interpretability method because it describes three desirable properties as follows: (i) local accuracy, (ii) missingness and (iii) consistency.^{25,30} To show how the ML model works, we used the SHAP algorithm to explain the individual diagnosis of several representative cases from our external validation dataset. The tree-based SHAP algorithm from the Python “shap” package by Lundberg was applied for the ML model.³¹

In particular, we used SHAP feature importance to rank features by reducing their importance as a measure of the average absolute Shapley values.³² Features with larger absolute SHAP value are deemed to be more important. Although the importance plot is useful, it does not contain information beyond the importance. The summary plot can reflect both feature importance and feature effects.³³ Each

TABLE 2 Performance comparison of ML models and ACR TI-RADS in the training dataset

Models	AUC	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)
US-only						
ACR TI-RADS	0.862	91.8	67.9	51.4	95.7	74.4
	(0.847-0.876)	(538/586)	(1074/1582)	(538/1046)	(1074/1122)	(1612/2168)
		[89.4-93.8]	[65.6-70.2]	[49.6-53.4]	[94.5-96.7]	[72.6-76.0]
RF	0.935	91.3	83.4	67.0	96.3	85.5
	(0.924-0.945)	(535/586)	(1319/1582)	(535/798)	(1319/1370)	(1854/2168)
		[88.9-93.5]	[81.5-85.1]	[64.6-69.6]	[95.3-97.2]	[84.0-86.9]
SVM	0.885	91.3	73.3	55.9	95.8	78.1
	(0.871-0.899)	(535/586)	(1159/1582)	(535/958)	(1159/1210)	(1694/2168)
		[89.1-93.3]	[71.0-75.3]	[53.8-58.0]	[94.7-96.8]	[76.4-79.8]
XGBoost	0.919	91.3	81.5	64.7	96.2	84.2
	(0.907-0.930)	(535/586)	(1290/1582)	(535/827)	(1290/1341)	(1825/2168)
		[88.9-93.5]	[79.5-83.4]	[62.1-67.2]	[95.2-97.1]	[82.6-85.7]
US-clinical						
RF	0.935	91.1	81.1	64.1	96.1	83.8
	(0.923-0.946)	(534/586)	(1283/1582)	(534/833)	(1283/1335)	(1817/2168)
		[88.7-93.2]	[79.1-83.0]	[61.8-66.5]	[95.1-97.0]	[82.3-85.3]
SVM	0.893	91.1	73.7	56.2	95.7	78.4
	(0.878-0.908)	(534/586)	(1166/1582)	(534/950)	(1166/1218)	(1700/2168)
		[88.7-93.3]	[71.4-76.0]	[54.0-58.4]	[94.7-96.8]	[76.6-80.1]
XGBoost	0.902	91.3	72.8	55.4	95.8	77.8
	(0.887-0.917)	(535/586)	(1152/1582)	(535/965)	(1152/1203)	(1687/2168)
		[88.9-93.5]	[70.7-75.0]	[53.5-57.6]	[94.7-96.8]	[76.2-79.5]

Note: Data in parentheses are 95% confidence intervals (CIs) or proportions.

Abbreviations: ACR TI-RADS, American College of Radiology Thyroid Imaging Reporting and Data System; AUC, area under the receiver operating characteristic curve; NPV, negative predictive value; PPV, positive predictive value; RF, Random Forest; SVM, Support Vector Machine; US, ultrasonography; XGBoost, extreme gradient boosting.

point on the summary plot denotes a SHAP value per feature and an instance. The vertical axis is the features ordered by their relative importance, while the horizontal axis is the SHAP value that represents the feature impact on model output. Thereby, the summary plot shows us the relationship between the feature values and the impact on the diagnosis.

The SHAP interpretation force plots can be used to visualize feature attributes such as the SHAP value as “forces,” which likes an arrow that pushes to increase (shown in red) or decrease (shown in blue) the probability of malignancy from the baseline.²³ These “forces” balance each other out at the actual diagnosis of the data instance. For global explanations for an entire dataset, we rotated many force plots vertically and placed them side by side on the basis of their clustering similarity.²³

2.6 | Statistical analysis

ML models were developed using the Python module H2O.ai. (2020, Python Interface version 3.30.0.6) and scikit-learn³⁴ (2021, Python Interface version 0.24.1). Statistical analyses were conducted using

R version 4.0.3 (<http://www.Rproject.org>). To evaluate the performance of ACR TI-RADS and ML models, six metrics of AUC, sensitivity, specificity, accuracy, positive predictive value (PPV) and negative predictive value (NPV) were calculated using the R package “pROC.”³⁵ Delong's test was used to compare the AUCs between two models. The 95% confidence interval (CI) was obtained by bootstrapping with 2000 samples. The optimal cutoff value was identified using the maximal Youden index. A *P*-value <.05 is considered statistically significant.

3 | RESULTS

3.1 | Patient and nodule characteristics

A total of 2880/3965 (72.6%) thyroid nodules were benign and 1085 (27.4%) were malignant. The entire cohort consisted of a training dataset (*n* = 2168), an internal validation dataset (*n* = 930) and an external validation dataset (*n* = 867). The mean age of patients was 46.1 years ±12.6 for the training dataset, 46.5 years ±13.2 for the internal validation dataset and 47.5 years ±13.0 for the external validation dataset (Table 1). The rate of malignancy in the three datasets

TABLE 3 Performance comparison of ML models and ACR TI-RADS in the internal validation dataset

Models	AUC	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)
US-only						
ACR TI-RADS	0.857 (0.832-0.878)	91.2 (241/264) [89.2-93.8]	68.3 (455/666) [65.5-70.2]	51.4 (241/452) [49.6-53.5]	95.7 (455/478) [94.5-96.7]	74.8 (696/930) [72.1-77.6]
RF	0.916 (0.896-0.933)	91.7 (242/264) [88.3-94.7]	78.7 (524/666) [75.7-81.7]	63.0 (242/384) [59.7-66.5]	96.0 (524/546) [94.4-97.5]	82.4 (766/930) [80.0-84.7]
SVM	0.869 (0.846-0.890)	89.4 (236/264) [85.6-92.8]	76.6 (509/666) [73.1-79.6]	60.1 (236/393) [56.7-63.7]	94.8 (509/537) [93.1-96.5]	80.1 (745/930) [77.5-82.6]
XGBoost	0.904 (0.883-0.922)	91.7 (242/264) [88.3-94.7]	78.1 (520/666) [74.9-81.2]	62.4 (242/388) [59.0-65.9]	96.0 (520/542) [94.4-97.4]	81.9 (762/930) [79.5-84.3]
US-clinical						
RF	0.914 (0.894-0.935)	91.7 (242/264) [88.3-94.7]	78.8 (524/666) [75.7-81.8]	63.1 (242/384) [59.8-66.8]	96 (524/546) [94.4-97.4]	82.5 (766/930) [80.0-84.8]
SVM	0.877 (0.852-0.901)	91.7 (242/264) [88.3-95.1]	73.3 (488/666) [69.8-76.7]	57.6 (242/420) [54.4-61.1]	95.7 (488/510) [93.9-97.3]	78.5 (730/930) [75.8-81.2]
XGBoost	0.887 (0.862-0.912)	91.7 (242/264) [88.3-95.1]	66.7 (444/666) [63.1-70.3]	52.2 (242/464) [49.4-55.2]	95.3 (444/466) [93.4-97.1]	73.9 (686/930) [71.0-76.6]

Note: Data in parentheses are 95% confidence intervals (CIs) or proportions.

Abbreviations: ACR TI-RADS, American College of Radiology Thyroid Imaging Reporting and Data System; AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value; RF, random forest; SVM, support vector machine; US, ultrasonography; XGBoost, extreme gradient boosting.

was 27.1% (586/2168), 28.3% (264/930) and 27.2% (235/867), respectively. The distribution of five ACR TI-RADS categories and nodule size between benign and malignant nodules in the training and validation datasets was shown in Table S2. Table S3 shows that age, underlying thyroid diseases, T3, T4, TT3, TT4 and TSH were identified as independent clinical variables for malignancy risk. As a result, a total of 14 features were included in the US-clinical ML models.

3.2 | Comparison of diagnostic performance between ML models and ACR TI-RADS

Tables 2-4 and Figure 2 demonstrate the diagnostic performance of ACR TI-RADS and ML models in the training, internal validation and external validation datasets on a lesion-level analysis. The optimal cut-off value for determining malignant nodules using ACR TI-RADS was TR4. In the training dataset, the ACR TI-RADS achieved an AUC of 0.862 (95% CI: 0.847-0.876), with a sensitivity of 91.8% (95% CI: 89.4-93.8), a specificity of 67.9% (95% CI: 65.6-70.2) and an accuracy of 74.4% (95% CI: 72.6-76.0). All US-only ML models outperformed

ACR TI-RADS ($P < .001$), especially the RF model (hereafter, Thy-Wise), with an AUC of 0.935 (95% CI: 0.924-0.945), sensitivity of 91.3% (95% CI: 88.9-93.5), specificity of 83.4% (95% CI: 81.5-85.1) and accuracy of 85.5% (95% CI: 84.0-86.9). A web-based malignancy risk estimator for thyroid nodules using Thy-Wise is available at <https://www.xsmartanalysis.com/MedAI-Lab/Thy-Wise>.

Similarly, in the internal validation dataset, the AUC of Thy-Wise was higher than that of ACR TI-RADS (0.916 [95% CI: 0.896-0.933] vs 0.857 [95% CI: 0.832-0.878]; $P < .001$). Their sensitivities were similar (91.7% [95% CI: 88.3-94.7] vs 91.2% [95% CI: 89.2-93.8]), whereas the specificity of Thy-Wise was higher than that of ACR TI-RADS (78.7% [95% CI: 75.7-81.7] vs 68.3% [95% CI: 65.5-70.2]; $P < .001$).

We evaluated the generalization ability of the ACR TI-RADS and ML models on an external validation dataset. The Thy-Wise had higher AUC than did ACR TI-RADS (0.904 [95% CI: 0.883-0.923] vs 0.853 [95% CI: 0.828-0.876]; $P < .001$). The specificity of Thy-Wise was higher than that of ACR TI-RADS (78.5% [95% CI: 75.2-81.8] vs 66.9% [95% CI: 63.4-70.2]). The accuracy of Thy-Wise was 82.1% (95% CI: 79.6-84.7), which was higher than that of ACR TI-RADS

TABLE 4 Performance comparison of ML models and ACR TI-RADS in the external validation dataset

Models	AUC	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)
US-only						
ACR TI-RADS	0.853	91.1	66.9	50.6	95.2	73.4
	(0.828-0.876)	(214/235)	(423/632)	(214/423)	(423/444)	(637/867)
		[66.3-94.4]	[63.4-92.2]	[47.9-75.0]	[88.2-97.0]	[70.8-84.0]
RF	0.904	91.9	78.5	61.3	96.3	82.1
	(0.883-0.923)	(216/235)	(496/632)	(216/352)	(496/515)	(712/867)
		[88.1-95.3]	[75.2-81.8]	[57.8-65.2]	[94.7-97.8]	[79.6-84.7]
SVM	0.864	90.2	58.9	44.9	94.2	67.4
	(0.840-0.887)	(212/235)	(372/632)	(212/472)	(372/395)	(584/867)
		[86.4-93.6]	[54.9-62.7]	[42.4-47.6]	[92.0-96.3]	[64.2-70.4]
XGBoost	0.865	87.7	64.4	47.8	93.4	70.7
	(0.840-0.887)	(206/235)	(407/632)	(206/431)	(407/436)	(613/867)
		[83.4-91.9]	[60.6-68.0]	[45.0-50.8]	[91.1-95.5]	[67.7-73.7]
US-clinical						
RF	0.880	91.1	63.8	48.3	95	71.2
	(0.855-0.905)	(214/235)	(403/632)	(214/443)	(403/424)	(617/867)
		[87.2-94.5]	[60.0-67.4]	[45.7-51.1]	[93.0-96.9]	[68.3-73.9]
SVM	0.854	90.2	57.6	44.1	94.1	66.4
	(0.825-0.884)	(212/235)	(364/632)	(212/480)	(364/387)	(576/867)
		[86.4-93.6]	[53.8-61.7]	[41.7-46.8]	[91.9-96.2]	[63.4-69.6]
XGBoost	0.864	87.7	66.9	49.6	93.7	72.5
	(0.838-0.891)	(206/235)	(423/632)	(206/415)	(423/452)	(629/867)
		[83.4-91.9]	[63.3-70.7]	[46.8-52.9]	[91.5-95.7]	[69.8-75.5]

Note: Data in parentheses are 95% confidence intervals (CIs) or proportions.

Abbreviations: ACR TI-RADS, American College of Radiology Thyroid Imaging Reporting and Data System; AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value; RF, random forest; SVM, support vector machine; US, ultrasonography; XGBoost, extreme gradient boosting.

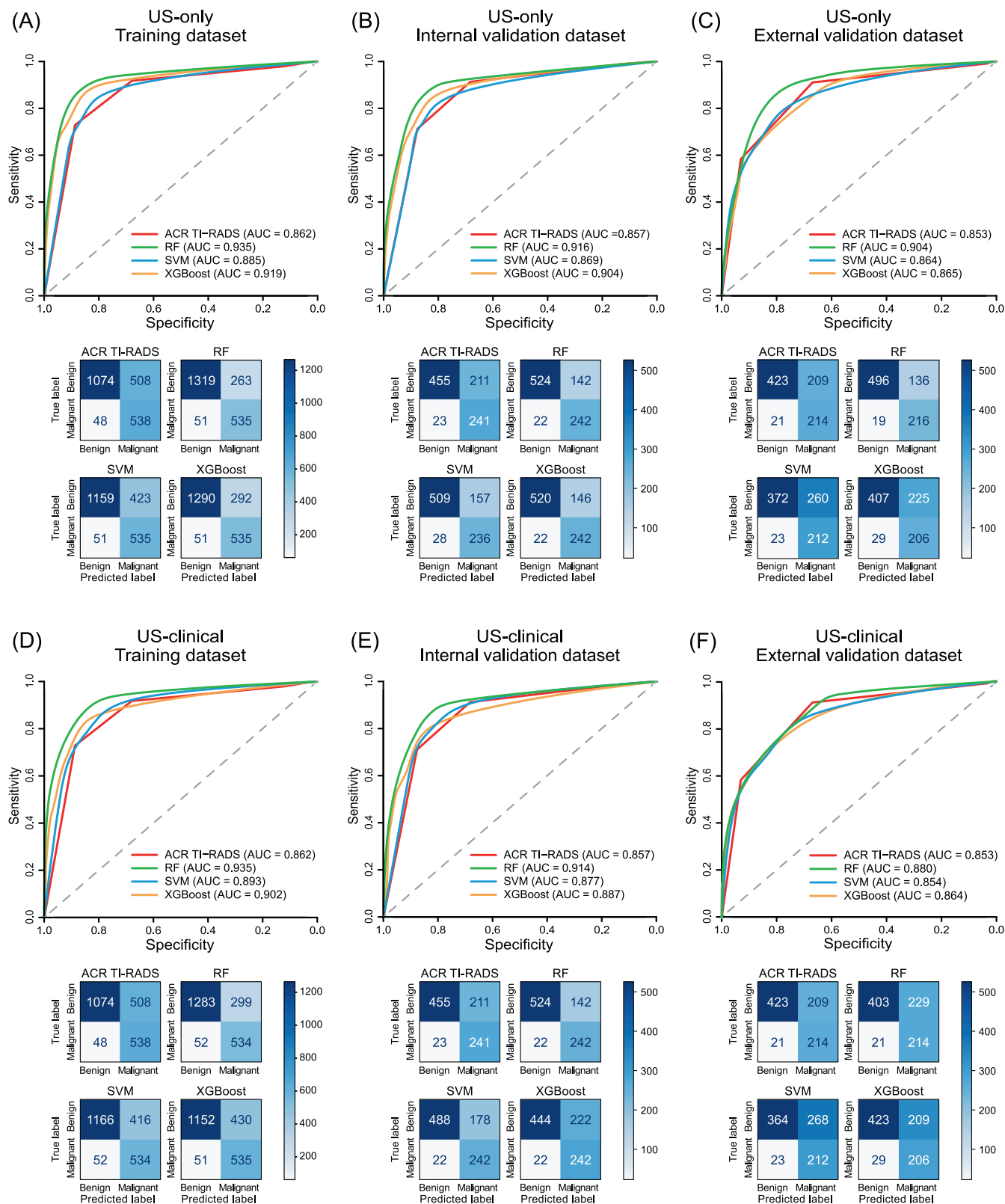


FIGURE 2 Diagnostic performance of ACR TI-RADS and ML models in discrimination of malignant from benign thyroid nodules. ROC curves and corresponding confusion matrices of US-only ML models compared to ACR TI-RADS in the training (A), internal validation (B) and external validation datasets (C). ROC curves and confusion matrices of US-clinical ML models compared to ACR TI-RADS in the training (D), internal validation (E) and external validation datasets (F). ACR TI-RADS, American College of Radiology Thyroid Imaging, Reporting and Data System; AUC, area under the curve; ML, machine learning; RF, random forest; ROC, receiver operator characteristic; SVM, support vector machine; XGBoost, extreme gradient boosting

(73.4%, 95% CI: 70.8-84.0). The sensitivity of Thy-Wise was comparable to that of ACR TI-RADS (91.9% [95% CI: 88.1-95.3] vs 91.1% [95% CI: 66.3-94.4]).

However, the addition of clinical data to the gray-scale US features showed no improvement in the evaluation of thyroid nodules (Table S4).

3.3 | Comparison of the unnecessary FNAB rate between ML models and ACR TI-RADS

At a similar biopsy rate of malignancy, the Thy-Wise had a lower unnecessary FNAB rate than did ACR TI-RADS recommendation (15.3% vs 32.3% for the internal validation dataset, 15.7% vs 47.3% for the external validation dataset) and US-only SVM (16.9% for the internal validation dataset and 30.0% for the external validation dataset), as well as US-only XGBoost (15.7% for the internal validation

dataset and 26.0% for the external validation dataset) (Tables 5, 6 and Figure 3). Adding clinical data to the US-only models (ie, US-clinical models) could not further reduce the unnecessary FNAB rate (Tables 5 and 6).

3.4 | Interpretability of Thy-Wise

The SHapley Additive exPlanation (SHAP) feature importance plots exhibit the margin and echogenic foci were the two most important features in the Thy-Wise, followed by shape, echogenicity, nodule size and composition (Figure 4). The plot depicts the distribution of the low to high Shapley values (impact on model output) for each feature. The higher the SHAP value of a feature, the greater probability of cancer. In the summary plot, we could observe the association between a feature's SHAP value and its impact on the diagnosis. Figure 5 displays the SHAP explanation force plots for four representative cases from

TABLE 5 Comparison of reliability for recommending FNAB on the internal validation dataset

Models	No. of recommended FNABs	No. of malignant nodules	No. of benign nodules	Biopsy rate of malignancy (%)	Unnecessary FNAB rates (%)
US-only					
ACR TI-RADS	525	225	300	85.2 (225/264)	32.3 (300/930)
RF	384	242	142	91.7 (242/264)	15.3 (142/930)
SVM	393	236	157	89.4 (236/264)	16.9 (157/930)
XGBoost	388	242	146	91.7 (242/264)	15.7 (146/930)
US-clinical					
RF	384	242	142	91.7 (242/264)	15.3 (142/930)
SVM	420	242	178	91.7 (242/264)	19.1 (178/930)
XGBoost	464	242	222	91.7 (242/264)	23.9 (222/930)

Note: Data in parentheses are proportions.

Abbreviations: ACR TI-RADS, American College of Radiology Thyroid Imaging Reporting and Data System; FNAB, fine-needle aspiration biopsy; RF, random forest; SVM, support vector machine; US, ultrasonography; XGBoost = extreme gradient boosting.

TABLE 6 Comparison of reliability for recommending FNAB on the external validation dataset

Models	No. of recommended FNABs	No. of malignant nodules	No. of benign nodules	Biopsy rate of malignancy (%)	Unnecessary FNAB rates (%)
US-only					
ACR TI-RADS	615	205	410	87.2 (205/235)	47.3 (410/867)
RF	352	216	136	91.9 (216/235)	15.7 (136/867)
SVM	472	212	260	90.2 (212/235)	30.0 (260/867)
XGBoost	431	206	225	87.7 (206/235)	26.0 (225/867)
US-clinical					
RF	443	214	229	91.1 (214/235)	26.4 (229/867)
SVM	480	212	268	90.2 (212/235)	30.9 (268/867)
XGBoost	415	206	209	87.7 (206/235)	24.1 (209/867)

Note: Data in parentheses are proportions.

Abbreviations: ACR TI-RADS, American College of Radiology Thyroid Imaging Reporting and Data System; FNAB, fine-needle aspiration biopsy; RF, random forest; SVM, support vector machine; US, ultrasonography; XGBoost, extreme gradient boosting.

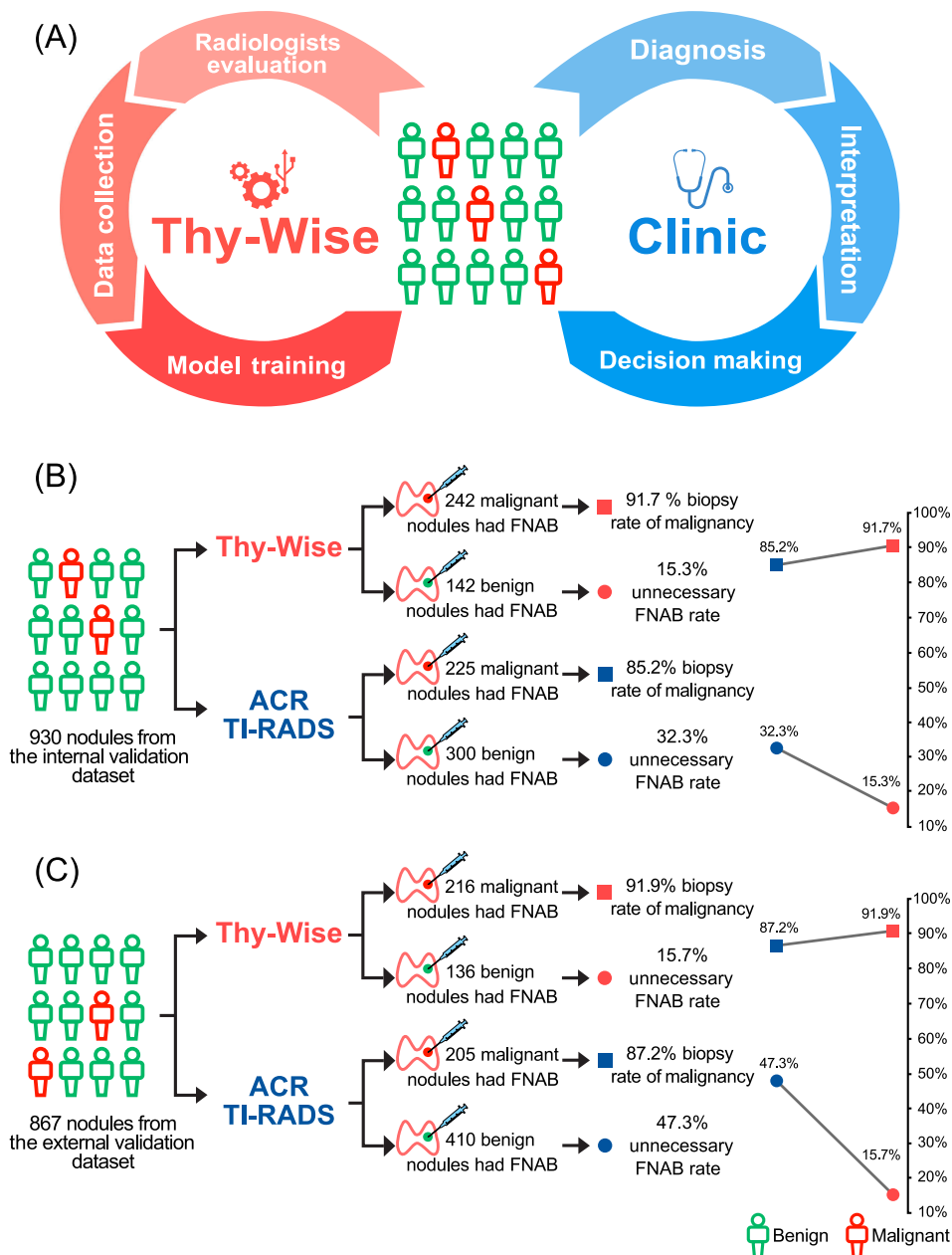


FIGURE 3 The construction of Thy-Wise and clinical management of thyroid nodules via Thy-Wise. (A) Workflow of the construction of Thy-Wise and its potential clinical application. (B) Comparison of Thy-Wise and ACR TI-RADS in FNAB recommendation on the internal validation dataset. The Thy-Wise assisted strategy reduced unnecessary FNAB rates from 32.3% to 15.3%, while achieving a higher biopsy rate of malignancy than ACR TI-RADS (85.2% vs 91.7%). (C) Comparison of Thy-Wise and ACR TI-RADS in FNAB recommendation on the external validation dataset. The Thy-Wise assisted strategy reduced unnecessary FNAB rates from 47.3% to 15.7%, while reaching a higher biopsy rate of malignancy than ACR TI-RADS (87.2% vs 91.9%). ACR TI-RADS, American College of Radiology Thyroid Imaging, Reporting and Data System; FNAB, fine-needle aspiration biopsy

the external validation dataset, which demonstrates how the US features push the output of the model from the baseline. The examples show Thy-Wise not only could report an exact probability of malignancy but also could provide a better indication for FNAB. Figure 6 shows SHAP supervised clustering by explanation similarity. Red SHAP values would increase the probability of malignancy, whereas blue values decrease it. A cluster that stands out on the left is a group with a high probability of malignancy.

4 | DISCUSSION

To the best of our knowledge, this large-scale retrospective study is the first to evaluate an interpretable ML strategy as an auxiliary tool for thyroid nodule management. Our Thy-Wise not only improved the

accuracy and specificity at similar sensitivity of evaluating thyroid nodules but also reduced the unnecessary FNAB rate as compared to ACR TI-RADS scoring system. Our simple ML model was developed based on human interpretation and their outputs also could be explained using the SHAP algorithm. The addition of clinical data to the gray-scale US could not provide additional diagnostic information.

From a clinical point of view, a fully automatic thyroid nodule diagnosis tool would benefit further management of patients, which is valuable and desperately needed. Many attempts have been made to improve the diagnostic performance via ML approaches. The computer-aided diagnosis (CAD) systems have been built to overcome the shortcomings of US diagnosis by radiologists, ultimately resulting in a reduction in unnecessary FNAB.^{36,37} A growing body of evidence suggests the efficacy of these systems in the diagnosis of thyroid nodules and has shown comparable or even superior performance to

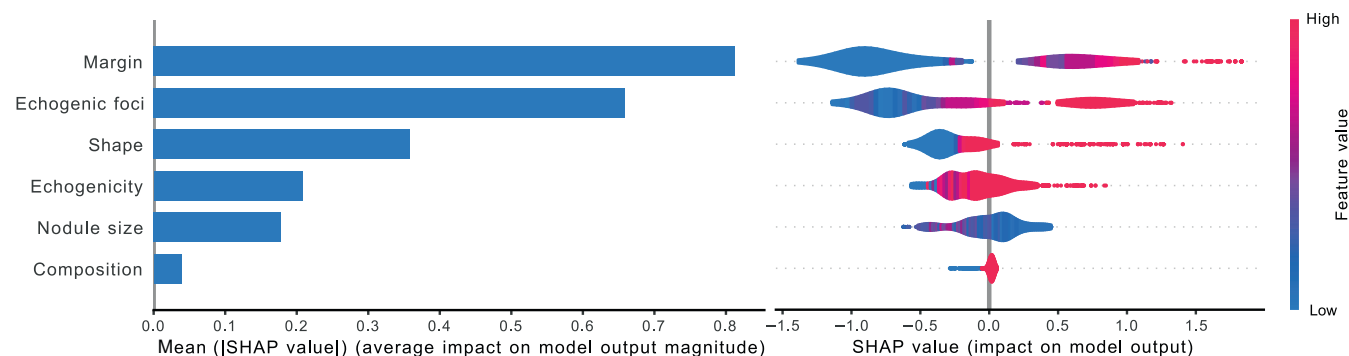


FIGURE 4 SHAP feature importance and summary plots for Thy-Wise. Standard bar charts (left panel) show that the features were ranked by decreasing importance measured as the mean absolute Shapley values. The larger is the average absolute SHAP value of the feature, the more important the feature to the model. The SHAP summary plots (right panel) show the distribution of the effect of each feature on the model output. Each dot is created for each feature attribution value of each nodule, multiple dots of the same feature attribution value converge into violin plots, using width to represent nodule density. Red represents higher feature encoding values, and blue represents lower, and intermediate values are depicted in purple. The value of the horizontal axis represents the effect of SHAP values on the model output, and the larger the value, the greater the probability of malignancy. SHAP, SHapley Additive exPlanation; T3, free triiodothyronine; T4, free thyroxine; TSH, thyrotropin; TT3, total triiodothyronine; TT4, total thyroxine; US, ultrasonography

experts or current risk stratification systems. Zhang et al demonstrated that ML classifiers were superior to experienced radiologists in the classification of thyroid nodules.¹⁹ Zhao et al found that the ML model significantly outperformed ACR TI-RADS in terms of classification performance and unnecessary FNAB rate.²⁶ Li et al newly developed a DL model that had higher specificity (86.1%-87.8% vs 57.1%-68.6%) than skilled radiologists at similar sensitivity (84.3%-93.4% vs 89.0%-96.9%) for thyroid nodule diagnosis in a real-world setting.²¹ Peng et al trained and validated a DL-based ThyNet using US images to differentiate thyroid nodules.²² The ThyNet-aided strategy not only outperformed radiologists but also could improve the diagnostic accuracy of radiologists and help decrease the unnecessary FNAB rate from 61.9% to 35.2%.²² However, the diagnostic accuracy of radiologists may be underestimated by following strict diagnostic criteria and reviewing static US images. It would be of great clinical significance to develop a DL model based on dynamic videos rather than static images and then compared the diagnostic performance with radiologists in a real-world setting. A recent meta-analysis showed that experienced radiologists may have an advantage over ML systems during real-time diagnosis.³⁷ In our work, we compared our human knowledge-based ML models with ACR TI-RADS, the most widely accepted guideline for thyroid nodule management. Thy-Wise outperformed ACR TI-RADS in accuracy (82.1% vs 73.4%), specificity (78.5% vs 66.9%) and unnecessary FNAB rates (15.7% vs 47.3%).

Most importantly, the intrinsic black-box nature of previous ML models for thyroid nodules has become a serious obstacle to explain their outputs. We are unable to understand how the US features lead to a diagnosis. To date, no studies have tried a strategy or methodology to directly explained the models. These models were less actionable due to a lack of transparency and interpretability, which would limit their applications in clinical practice.³⁸ With the rapid iteration of ML algorithms in recent years, the accuracy of ML models has

gradually reached a bottleneck, yet the challenge of model interpretability remains unresolved.¹⁹⁻²² Since the prediction process of ML models is complex and the decision basis is difficult to quantify, the results of the models cannot be understood intuitively through mathematical formulas as in the case of generalized linear models such as logistic regression.²⁴ There are calls for agnostic methods to assist the interpretation of ML models regardless of their complexity in a non-linear manner. However, clinical practice that involves complex ethical and moral issues requires careful decision-making, rendering evidence-based medicine an indispensable cornerstone of current medicine.³⁸ The prediction process will need to be transparent to ensure medical professionals embrace this new technology.

In our study, we leveraged a novel tree-based SHAP approach to visually explain the output of the RF model, which had the most outstanding diagnostic performance. RF is an ensemble learning algorithm that uses bootstrap aggregating (bagging) method to combine multiple decision tree models.²⁵ Decision tree is a classifier that keeps dividing the input data into two portions at each node of the tree until all the data is divided into single instances. RF applies the bagging method to randomly sample from the training dataset to build several different decision tree models, and finally combine the results of all decision trees by voting to derive the final output.³⁹ Consequently, RF inherits the easy-to-understand nature of the decision tree and thus can be interpreted by tree-based SHAP approach, and the occurrence of overfitting is effectively avoided by the bagging method, resulting in superior performance.²⁵ Shapley values, on the other hand, can be used to estimate the contribution magnitude (ie, feature importance) and direction of features.⁴⁰ Features with a positive direction pull the diagnosis toward malignancy, whereas features with a negative direction pull the diagnosis toward benignity. The SHAP provides intuitive visualizations of feature contributions, which helps the perception of black-box in ML.²³ The explanation informs the clinicians which elements in the current ML model are potentially actionable, thus



FIGURE 5 The SHAP local force plots for four representative cases using Thy-Wise. The SHAP local force plot provides individual-level interpretability. The SHAP value for each feature represents as a force, which either increases (red) or decreases (blue) the predicted probability of malignancy from the baseline. Nodules with an estimated probability of malignancy above the threshold will be classified as malignant and recommended to undergo FNAB. (A) Nodule 1 was classified as TR4 with a maximum diameter of 22 mm, thus, ACR TI-RADS recommended FNAB. However, the biopsy result suggested benignity. SHAP force plot shows four US features (hypoechoic, macrocalcifications, ill-defined and nodule size) push the diagnosis toward malignancy while other two US features (mixed cystic and solid and wider-than-tall) push the diagnosis toward benignity. Ultimately, Thy-Wise evaluated the malignancy probability of this nodule as 31% and yields benign predictions, avoiding the nodule from unnecessary biopsy. (B) Nodule 2 was classified as TR4 with a maximum diameter of 14 mm, ACR TI-RADS recommended no need for FNAB. However, SHAP force plot shows four US features (mixed cystic and solid, ill-defined, punctate echogenic foci and nodule size) push the diagnosis toward malignancy, only echogenicity pushes the diagnosis toward benignity. Thy-Wise reported a malignancy rate of 70% and recommended FNAB. The pathological finding confirmed malignancy, suggesting that Thy-Wise could facilitate timely intervention for similar malignant nodules. (C) Nodule 3 was classified as TR5 with a maximum diameter of 13 mm, ACR TI-RADS recommended FNAB. SHAP force plot shows only echogenic foci mildly pushes the diagnosis toward benignity despite the others pushing the diagnosis toward malignancy. Thus Thy-Wise assessed the malignancy rate of this nodule as 81% and recommended FNAB, establishing the basis for personalized intervention for thyroid nodules. The pathological finding also showed malignancy. (D) Nodule 4 was classified as TR3 with a maximum diameter of 20 mm, ACR TI-RADS thus recommended no need for FNAB. SHAP force plot shows three US features (hypoechoic, macrocalcifications and nodule size) push the diagnosis toward malignancy whereas the spongiform pushes the diagnosis toward benignity. Finally, Thy-Wise assessed the malignancy rate of this nodule as 11% and recommended no need for FNAB. The biopsy result confirmed benignity. ACR TI-RADS, American College of Radiology Thyroid Imaging Reporting and Data System; FNAB, fine-needle aspiration biopsy; SHAP, SHapley Additive exPlanation; TR, thyroid imaging reporting and data system risk level

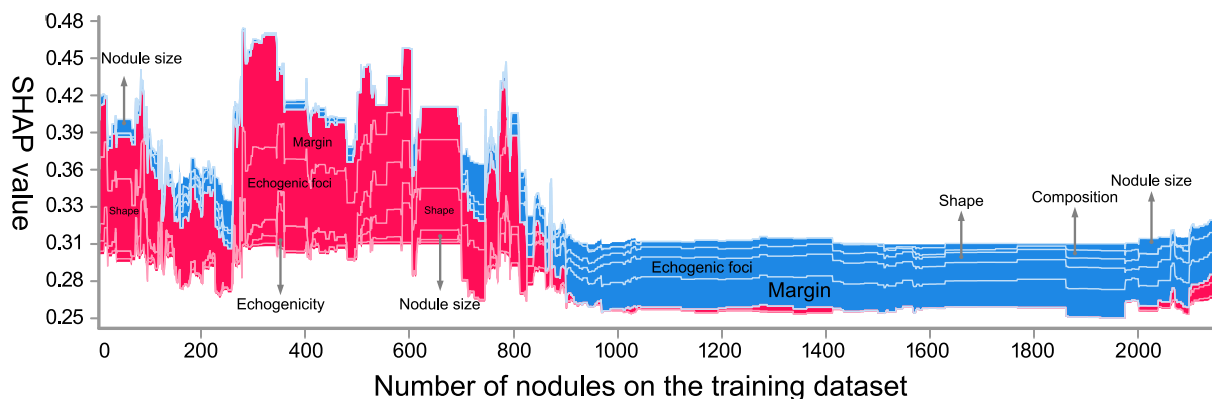


FIGURE 6 The SHAP global force plots for Thy-Wise. The SHAP global force plot is an integration of the local force plots for each individual in an entire dataset, representing the feature importance at both global and individual levels to provide interpretability to the inference process of the model. Each band denotes a feature and its width represents importance; red means the category of the feature increases the predicted probability of malignancy, while blue means that the category of the feature decreases the probability of malignancy. SHAP, SHapley Additive exPlanation; T3, free triiodothyronine; T4, free thyroxine; TSH, thyrotropin; TT3, total triiodothyronine; TT4, total thyroxine; US, ultrasonography

rendering the model appropriate for a further test as a clinical decision marking tool.

We also explored the incremental value of clinical characteristics to the gray-scale US for the diagnosis of thyroid nodules. Given that thyroid cancer is complex and heterogeneous, multiresource data in practical clinical scenarios, such as demographics, laboratory findings (eg, thyroid function tests) and US images, are encouraged to combine to create a stronger ML model.⁴¹ Thyroid function tests are commonly involved in the initial evaluation of thyroid nodules and the relationship between thyroid function and the risk of thyroid cancer has been investigated. TSH concentration could be used as a biomarker for the prediction of thyroid malignancy.⁴² T4 was elevated in thyroid cancer⁴³ whereas the inverse has been shown for TT3.⁴⁴ Moreover, mutated forms of T3 receptor were found may increase phosphatidylinositol 3-kinase signaling and induce thyroid carcinomas in a mouse model.⁴⁵ In our study, although we observed that thyroid function tests were associated with nodule malignancy, they could not enhance the diagnostic performance of gray-scale US-based ML models in the classification of thyroid nodules.

Our study had several limitations. First, the retrospective nature may introduce recall bias. Second, an inevitable selection bias may exist because our study only included definitive benign or malignant thyroid nodules with biopsy cytological or surgical pathology results. Those benign nodules that did not undergo diagnostic FNAB were excluded. Third, small (<10 mm) papillary thyroid cancers were excluded given the increasing trend toward surveillance for this subgroup.⁵ Finally, we did not compare with other risk stratification systems because they have not been widely accepted.

5 | CONCLUSIONS

The application of online Thy-Wise may improve the diagnostic accuracy and specificity of evaluating thyroid nodules and reduce the

unnecessary FNAB rate as compared to ACR TI-RADS. We, for the first time, revealed that clinical data including thyroid function tests failed to improve the accuracy of ML models. We applied SHAP methods to explain the output of Thy-Wise, which may accelerate the interpretable model toward clinical application. Our study also encourages further use of the SHAP algorithm to better uncover ML efforts and improve model transparency.

AUTHOR CONTRIBUTIONS

Conception and design: Zhe Jin, Shufang Pei, Lizhu Ouyang, Bin Zhang and Shuixing Zhang. *Collection and assembly of data:* Shufang Pei, Lizhu Ouyang, Lu Zhang, Xiaokai Mo, Qiuying Chen, Luyan Chen and Jingjing You. *Development of methodology:* Zhe Jin and Bin Zhang. *Data analysis and interpretation:* Zhe Jin and Bin Zhang. *Manuscript writing:* Zhe Jin, Bin Zhang and Shuixing Zhang. *Final approval of manuscript:* All authors. The work reported in the paper has been performed by the authors, unless clearly specified in the text.

FUNDING INFORMATION

This research was supported by a grant of the National Natural Science Foundation of China (81571664), the Science and Technology Planning Project of Guangdong Province (2014A020212244, 2016A020216020), the Scientific Research General Project of Guangzhou Science Technology and Innovation Commission (201605110912158), the China Postdoctoral Science Foundation (2016M600145) and the Guangdong Grand Science and Technology Special Project (2015B010106008).

CONFLICT OF INTEREST

The authors declare that they have no conflict of interests.

DATA AVAILABILITY STATEMENT

Images are stored on the picture archiving and communication system (GE Centricity; GE Healthcare, Milwaukee, WI) and are available at

<https://doi.org/10.6084/m9.figshare.20417895.v1>. The corresponding code for model development has been uploaded to <https://github.com/MedAI-Lab/Thy-Wise>. Further information is available on request from the corresponding author.

ETHICS STATEMENT

This retrospective study was approved by the institutional Ethics Committees of the First Affiliated Hospital of Jinan University, which waived the requirement for informed patient consent due to the anonymity and retrospective nature.

ORCID

Shuixing Zhang  <https://orcid.org/0000-0001-7377-382X>

REFERENCES

- Guth S, Theune U, Aberle J, Galach A, Bamberger CM. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. *Eur J Clin Invest*. 2009;39:699-706.
- Durante C, Grani G, Lamartina L, Filetti S, Mandel SJ, Cooper DS. The diagnosis and management of thyroid nodules: a review. *JAMA*. 2018;319:914-924.
- Vaccarella S, Franceschi S, Bray F, Wild CP, Plummer M, Dal Maso L. Worldwide thyroid-cancer epidemic? The increasing impact of overdiagnosis. *N Engl J Med*. 2016;375:614-617.
- Ha EJ, Baek JH, Na DG. Risk stratification of thyroid nodules on ultrasonography: current status and perspectives. *Thyroid*. 2017;27:1463-1468.
- 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*. 2017;14:587-595.
- Tessler FN, Middleton WD, Grant EG. Thyroid imaging reporting and data system (TI-RADS): a user's guide. *Radiology*. 2018;287:29-36.
- Hoang JK, Middleton WD, Farjat AE, et al. Reduction in thyroid nodule biopsies and improved accuracy with American College of Radiology Thyroid Imaging Reporting and Data System. *Radiology*. 2018;287:185-193.
- Yang R, Zou X, Zeng H, Zhao Y, Ma X. Comparison of diagnostic performance of five different ultrasound TI-RADS classification guidelines for thyroid nodules. *Front Oncol*. 2020;10:598225.
- Castellana M, Castellana C, Treglia G, et al. Performance of five ultrasound risk stratification systems in selecting thyroid nodules for FNA. *J Clin Endocrinol Metab*. 2020;105:1659-1669.
- Li W, Wang Y, Wen J, Zhang L, Sun Y. Diagnostic performance of American College of Radiology TI-RADS: a systematic review and meta-analysis. *AJR Am J Roentgenol*. 2021;216:38-47.
- Zhang W-B, Xu H-X, Zhang Y-F, et al. Comparisons of ACR TI-RADS, ATA guidelines, Kwak TI-RADS, and KTA/KSThR guidelines in malignancy risk stratification of thyroid nodules. *Clin Hemorheol Microcirc*. 2020;75:219-232.
- Huang BL, Ebner SA, Makkar JS, et al. A multidisciplinary head-to-head comparison of American College of Radiology Thyroid Imaging and Reporting Data System and American Thyroid Association Ultrasound Risk Stratification Systems. *Oncologist*. 2020;25:398-403.
- Lauria Pantano A, Maddaloni E, Briganti SI, et al. Differences between ATA, AACE/ACE/AME and ACR TI-RADS ultrasound classifications performance in identifying cytological high-risk thyroid nodules. *Eur J Endocrinol*. 2018;178:595-603.
- Ha EJ, Na DG, Baek JH, Sung JY, Kim J-H, Kang SY. US fine-needle aspiration biopsy for thyroid malignancy: diagnostic performance of seven society guidelines applied to 2000 thyroid nodules. *Radiology*. 2018;287:893-900.
- Wildman-Tobriner B, Buda M, Hoang JK, et al. Using artificial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: diagnostic accuracy and utility. *Radiology*. 2019;292:112-119.
- Middleton WD, Teefey SA, Tessler FN, et al. Analysis of malignant thyroid nodules that do not meet ACR TI-RADS criteria for fine-needle aspiration. *AJR Am J Roentgenol*. 2021;216:471-478.
- Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol*. 2019;20:e262-e273.
- Ha EJ, Baek JH. Applications of machine learning and deep learning to thyroid imaging: where do we stand? *Ultrasonography*. 2021;40:23-29.
- Zhang B, Tian J, Pei S, et al. Machine learning-assisted system for thyroid nodule diagnosis. *Thyroid*. 2019;29:858-867.
- 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;9:17843.
- Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet Oncol*. 2019;20:193-201.
- Peng S, Liu Y, Lv W, et al. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multi-centre diagnostic study. *Lancet Digit Health*. 2021;3:e250-e259.
- Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2:749-760.
- Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Sydney*. NSW, Australia: Association for Computing Machinery; 2015:1721-1730.
- Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Vol XXII. New York, NY: Springer; 2009:745.
- Zhao C-K, Ren T-T, Yin Y-F, et al. A comparative analysis of two machine learning-based diagnostic patterns with thyroid imaging reporting and data system for thyroid nodules: diagnostic performance and unnecessary biopsy rate. *Thyroid*. 2021;31:470-481.
- Deshmukh F, Merchant SS. Explainable machine learning model for predicting GI bleed mortality in the intensive care unit. *Am J Gastroenterol*. 2020;115:1657-1668.
- Lundberg S, Lee S-I. *A Unified Approach to Interpreting Model Predictions*. Long Beach, CA: Neural Information Processing Systems (NIPS); 2017.
- Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst*. 2014;41:647-665.
- Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. Paper presented at: KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016:1135-1144.
- Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2:56-67.
- Bi Y, Xiang D, Ge Z, Li F, Jia C, Song J. An interpretable prediction model for identifying N7-Methylguanosine sites based on XGBoost and SHAP. *Mol Ther Nucleic Acids*. 2020;22:362-372.
- Tseng P-Y, Chen Y-T, Wang C-H, et al. Prediction of the development of acute kidney injury following cardiac surgery by machine learning. *Crit Care*. 2020;24:478.
- Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
- Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinf*. 2011;12:77.
- Zhao W-J, Fu L-R, Huang Z-M, Zhu J-Q, Ma B-Y. Effectiveness evaluation of computer-aided diagnosis system for the diagnosis of thyroid

- nodules on ultrasound: a systematic review and meta-analysis. *Medicine (Baltimore)*. 2019;98:e16379.
37. Xu L, Gao J, Wang Q, et al. Computer-aided diagnosis systems in diagnosing malignant thyroid nodules on ultrasonography: a systematic review and meta-analysis. *Eur Thyroid J*. 2020;9: 186-193.
 38. Haibe-Kains B, Adam GA, Hosny A, et al. Transparency and reproducibility in artificial intelligence. *Nature*. 2020;586:E14-E16.
 39. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods*. 2009;14: 323-348.
 40. Thorsen-Meyer H-C, Nielsen AB, Nielsen AP, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Digit Health*. 2020;2: e179-e191.
 41. Hu D, Peng F, Niu W. Deep convolutional neural network models for the diagnosis of thyroid cancer. *Lancet Oncol*. 2019;20:e129.
 42. Boi F, Minerba L, Lai ML, et al. Both thyroid autoimmunity and increased serum TSH are independent risk factors for malignancy in patients with thyroid nodules. *J Endocrinol Invest*. 2013;36: 313-320.
 43. Cho YA, Kong SY, Shin A, et al. Biomarkers of thyroid function and autoimmunity for predicting high-risk groups of thyroid cancer: a nested case-control study. *BMC Cancer*. 2014;14:873.
 44. Jonklaas J, Nsouli-Maktabi H, Soldin SJ. Endogenous thyrotropin and triiodothyronine concentrations in individuals with thyroid cancer. *Thyroid*. 2008;18:943-952.
 45. Furuya F, Ying H, Zhao L, Cheng SY. Novel functions of thyroid hormone receptor mutants: beyond nucleus-initiated transcription. *Steroids*. 2007;72:171-179.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Jin Z, Pei S, Ouyang L, et al. Thy-Wise: An interpretable machine learning model for the evaluation of thyroid nodules. *Int J Cancer*. 2022;151(12): 2229-2243. doi:[10.1002/ijc.34248](https://doi.org/10.1002/ijc.34248)