scientific reports

OPEN

ELTIRADS framework for thyroid nodule classification integrating elastography, TIRADS, and radiomics with interpretable machine learning

Erfan Barzegar-Golmoghani[1,10], Mobin Mohebi[1,9,10], Zahra Gohari[2,10], Sadaf Aram[1], Ali Mohammadzadeh[5], Sina Firouznia[6], Madjid Shakiba[3], Hamed Naghibi[3], Sadegh Moradian[7], Maryam Ahmadi[1], Kazhal Almasi[1], Mahbod Issaiy[3], Mehran Anjomrooz[2], Seyed Mohammad Tavangar[4], Sheida Javadi[3], Ahmad Bitarafan-Rajabi[8], Mohammad Davoodi[2✉], Hashem Sharifian[2✉] & Maryam Mohammadzadeh[2✉]

Early detection of malignant thyroid nodules is crucial for effective treatment, but traditional diagnostic methods face challenges such as variability in expert opinions and limited integration of advanced imaging techniques. This prospective cohort study investigates a novel multimodal approach, integrating traditional methods with advanced machine learning techniques. We studied 181 patients who underwent fine-needle aspiration (FNA) biopsy, each contributing one nodule, resulting in a total of 181 nodules for our analysis. Data collection included sex, age, and ultrasound imaging, which incorporated elastography. Features extracted from these images included Thyroid Imaging Reporting and Data System (TIRADS) scores, elastography parameters, and radiomic features. The pathological results based on the FNA biopsy, provided by the pathologists, served as our gold standard for nodule classification. Our methodology, termed ELTIRADS, combines these features with interpretable machine learning techniques. Performance evaluation showed that a Support Vector Machine (SVM) classifier using TIRADS, elastography data, and radiomic features achieved high accuracy (0.92), with sensitivity (0.89), specificity (0.94), precision (0.89), and F1 score (0.89). To enhance interpretability, we used hierarchical clustering, shapley additive explanations (SHAP), and partial dependence plots (PDP). This combined approach holds promise for enhancing the accuracy of thyroid nodule malignancy detection, thereby contributing to advancements in personalized and precision medicine in the field of thyroid cancer research.

Keywords Radiomics, Interpretable machine learning, Elastography, Nodule classification, Hierarchical clustering

Thyroid nodules, primarily benign and either solid or fluid-filled, are a common clinical issue. Epidemiological studies indicate that approximately 5% of women and 1% of men in iodine-sufficient regions have palpable thyroid nodules[1,2]. Most nodules are asymptomatic and typically found incidentally during routine examinations or imaging for unrelated conditions.

[1]Department of Biomedical Engineering, Tarbiat Modares University, Tehran, Iran. [2]Department of Radiology, Tehran University of Medical Science, Tehran, Iran. [3]Advanced Diagnostic and Interventional Radiology Research Center (ADIR), Tehran University of Medical Sciences, Tehran, Iran. [4]Department of Pathology, Tehran University of Medical Sciences, Tehran, Iran. [5]Department of Radiology, Iran University of Medical Sciences, Tehran, Iran. [6]Second Faculty of Medicine, Charles University, Prague, Czech Republic. [7]Department of Radiology, Kermanshah University of Medical Sciences, Kermanshah, Iran. [8]Rajaie Cardiovascular Medical and Research Institute, Iran University of Medical Sciences, Tehran, Iran. [9]Institut de Biologie Valrose (IBV), Université Côte d'Azur, CNRS, Inserm, Nice, France. [10]Erfan Barzegar-Golmoghani, Mobin Mohebi and Zahra Gohari: These authors contributed equally to this work. ✉email: mohammaddavoodi47@yahoo.com; hashemsharifian@gmail.com; m-mohammadzadeh@sina.tums.ac.ir

1

Although the majority of thyroid nodules are benign, some can grow large enough to cause symptoms like dysphagia or dyspnea. Additionally, certain nodules may overproduce thyroxine, leading to hyperthyroidism. Diagnosis relies heavily on clinical examination and imaging, with ultrasound showing a combined sensitivity of 0.88 [95% CI (0.83-0.91)] and specificity of 0.86 [95% CI (0.79-0.90)] for nodule detection[3].

Assessing malignancy risk often requires further investigation, as most thyroid cancers are slow-growing. While prognosis is generally favorable due to the benign nature of thyroid nodules[4], large nodules or multinodular goiters can lead to mechanical complications, and hormone-producing nodules can induce hyperthyroidism[5].

Elastography, an advanced imaging technique that measures tissue firmness, has gained prominence in evaluating thyroid nodules[6]. This method distinguishes between benign and malignant formations, enhancing differentiation accuracy beyond conventional ultrasound. Shear wave elastography (SWE) quantifies lesion stiffness, with tumors exceeding 65 kPa or a stiffness ratio greater than 3.7 considered highly suspicious[7]. This technique can detect malignant follicular tumors often overlooked by traditional fine needle aspiration (FNA) and can identify lymph node metastasis in papillary thyroid cancer due to increased stiffness[6–8].

The Thyroid Imaging Reporting and Data System (TIRADS) is a classification system for ultrasound imaging that aids clinicians in assessing malignancy risk and refining biopsy protocols[9]. The American College of Radiology's TIRADS utilizes a points-based system to evaluate nodules based on composition, echogenicity, shape, margin irregularities, and the presence of echogenic foci. Each feature is assigned a score, which determines the nodule's risk category and subsequent biopsy recommendations[10]. For nodules with indeterminate cytology, molecular testing can identify genetic markers for potential malignancy[11]. Combining this scoring system with artificial intelligence enhances diagnostic precision, as machine learning (ML) algorithms can efficiently analyze and interpret complex sonographic patterns[12].

The use of ML in medicine has rapidly increased, with many studies highlighting its role in enhancing diagnostic and prognostic capabilities[13,14]. Radiomics translates medical image information into quantitative attributes, facilitating detailed imaging data retrieval. These attributes improve treatment-related decisions and outcome predictions[15–17]. Previous investigations have shown that radiomic features derived from medical images are accurate, reliable, and non-invasive, particularly for cancer diagnosis, prognosis, and treatment outcome prediction[18,19]. Furthermore, integrating radiomics with ML offers a non-invasive path to discovering imaging biomarkers and developing predictive prognostic models[20–22].

Studies have demonstrated ML's potential in diagnosing thyroid nodules[23,24]. Yang et al. found that ML frameworks, particularly random forest algorithms, outperformed human judgment in predicting malignancy[25]. A comparative analysis of various ML algorithms for thyroid disease classification found that the random forest algorithm yielded the highest accuracy and specificity[26]. Zhang et al. proposed reducing subjectivity in thyroid nodule diagnosis by extracting geometric and morphological features from an ultrasound thyroid nodule image database using ML[27]. Another study developed a ML-based diagnostic model for thyroid nodules, showing that a random forest algorithm outperformed radiologists' diagnoses based on both conventional ultrasound and its combination with real-time elastography[28]. Wang et al. applied multimodal ultrasound radiomic technology, extracting features from 2D ultrasound, strain elastography (SE), and SWE images to predict the malignancy of thyroid nodules[29]. Additionally, another radiomics model, incorporating features from ultrasound and elastography, was used to predict malignancy in thyroid nodules[30]. These findings underscore ML's promising role in improving thyroid nodule diagnosis.

In light of these objectives, our study makes several key contributions to enhance the diagnosis of thyroid nodules:

- Integrated elastographic features with ultrasonography to improve assessment of nodule malignancy risk.
- Developed an SVM model that outperformed radiologists, serving as a valuable second-opinion tool.
- Highlighted the role of elastography in nodule classification through SHAP analysis.
- Identified atypical nodules via hierarchical clustering for further clinical investigation.
- Employed interpretable ML techniques to enhance prediction transparency.
- Established optimal thresholds for TIRADS and elastographic feature.

Materials and methods
Study design and participants
This prospective cohort study was conducted over one year at a tertiary teaching hospital's thyroid ultrasound clinic. We included patients with thyroid nodules who were eligible for FNA based on TIRADS criteria, specifically TIRADS 3 nodules measuring ≥ 25 mm, TIRADS 4 nodules measuring ≥ 15 mm, and TIRADS 5 nodules measuring ≥ 10 mm. Additionally, based on pathological criteria, patients with conclusive FNA results, including Bethesda category II and VI, were included. Patients were excluded if their pathology reports were inconclusive or indeterminate (Bethesda categories I and III), or if the aspirated specimen was inadequate. Participants were enrolled consecutively upon providing informed consent. Throughout the study, radiologists remained blinded to the final pathology results to ensure unbiased evaluation.

The Bethesda categories for thyroid cytology and their corresponding risk of malignancy (ROM) percentages are summarized in Table 1[31].

Data collection
Clinical and imaging data
Each patient underwent both ultrasonography and elastography of their thyroid nodule(s). Key features extracted from the nodules included echogenicity, shape, margins, internal echoes, vascularity, and elastographic parameters such as mean elasticity (Emean), maximum elasticity (Emax), minimum elasticity (Emin), and the Q-box ratio. The elastographic parameters were collected using the Supersonic AIXPLORER Mach 30

Category	ROM (%)
I: Nondiagnostic	-
II: Benign	< 3
III: Atypia of Undetermined Significance or Follicular Lesion of Undetermined Significance	10–30
IV: Follicular Neoplasm or Suspicious for Follicular Neoplasm	25–40
V: Suspicious for Malignancy	50–75
VI: Malignant	97–99

Table 1. Bethesda categories for thyroid cytology.
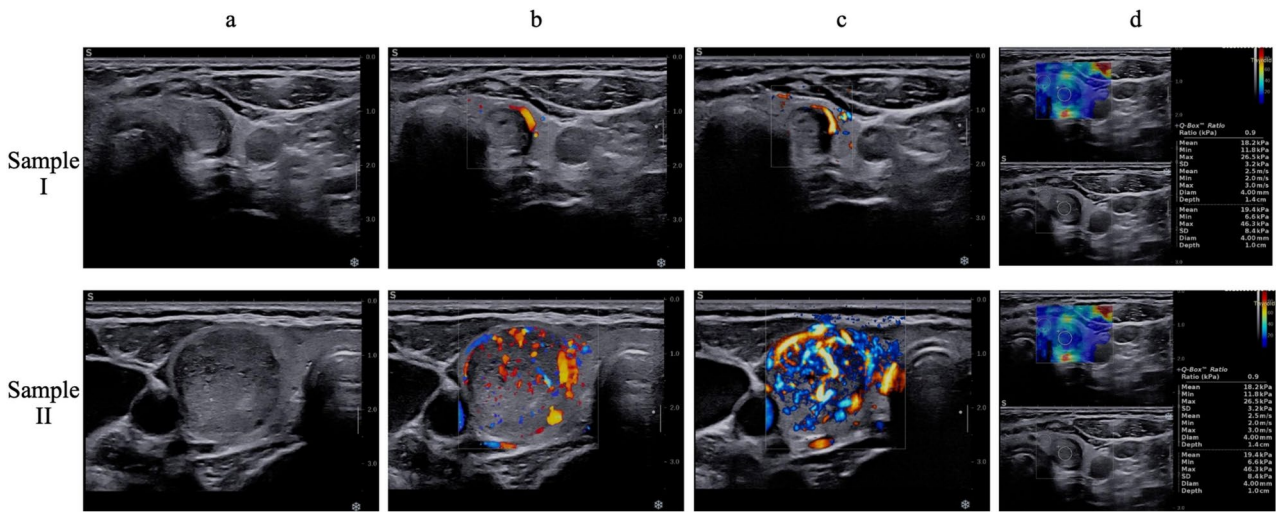
Fig. 1. It depicts two samples, labeled as I and II. Sample I is a thyroid nodule measuring 9 x 11 mm, characterized as hypoechoic and taller than wide (**a**). It exhibits peripheral vascularity in both color Doppler and Angio pl.u.s (**b**, **c**), and provides elastography parameters (**d**) for assessment. Sample II, in contrast, is a hypotrochoid thyroid nodule measuring 20 x 19 mm (**a**). This nodule features both punctate and coarse calcification. Additionally, it displays peripheral and central vascularity in color Doppler and Angio pl.u.s (**b**, **c**) and offers elastography parameters (**d**) for further evaluation.

system. Nodule characteristics were then assessed and scored according to TIRADS criteria. All nodules were independently evaluated by two radiologists with 3 and 5 years of experience, without knowledge of each other's diagnoses. In cases of disagreement, a more experienced radiologist with 10 years of experience provided the final diagnosis.

Subsequently, FNA was performed under ultrasonographic guidance. The aspirated specimens were analyzed by the pathology department and categorized as benign or malignant. Specimens that were inadequate or indeterminate were either repeated or excluded from the study.

Ethical considerations
This prospective study was approved by the Research Council of Tehran University of Medical Sciences (Ethical code: IR.TUMS.AMIRALAM.REC.1399.042). Participants were informed about the research objectives and given the option to participate, ensuring that their involvement did not interfere with their diagnosis or treatment. No additional costs were imposed on the patients. Confidentiality of all collected information was assured, and participants were informed that the results would be made available to them if desired. All procedures were conducted with the utmost respect for the rights and well-being of the participants, in compliance with relevant ethical standards.

Figures 1 and 2 illustrate representative thyroid nodule samples from our study. These images depict various characteristics observed through color Doppler, Angio pl.u.s, and elastographic imaging.

Sample size and power
Based on earlier research, we determined the appropriate sample size for thyroid nodules with a goal of a 5% α (type I error) and a power of 80% $(1 - \beta)$. Therefore, our sample size was set at 181.

The sample size calculation formula is given by:

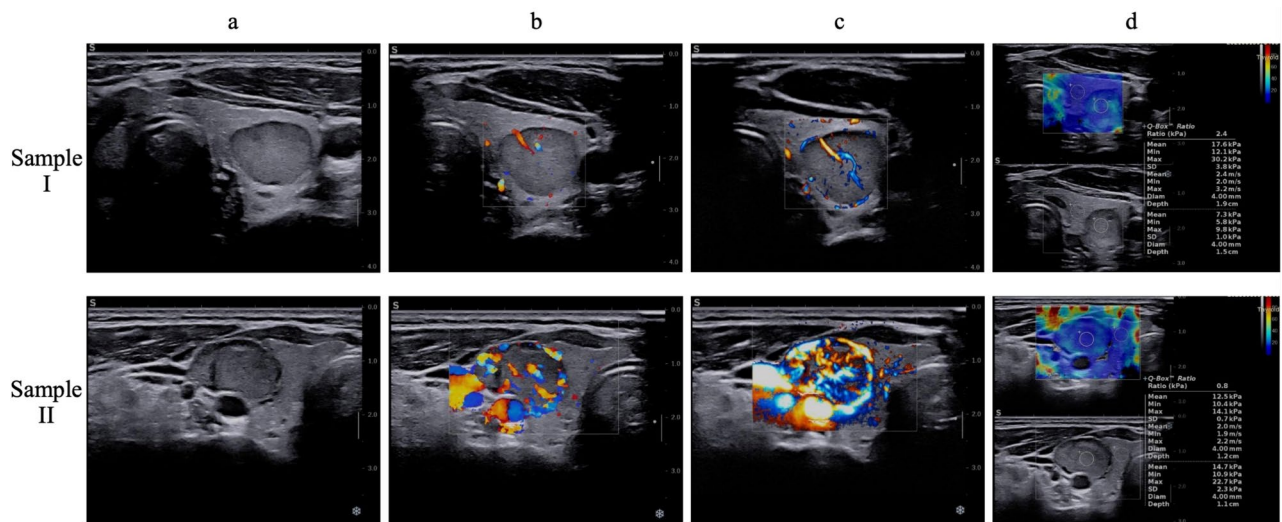$$n = \left(\frac{Z_{1-\alpha/2} \times \sigma}{d} \right)^2 \tag{1}$$

Fig. 2. It presents two samples, labeled as I and II. Sample I is a hypoechoic nodule with regular margins, measuring 15 x 12 mm (**a**). It displays peripheral vascularity on color Doppler (**b**) and both peripheral and central vascularity on Angio pl.u.s (**c**), along with elastography parameters (**d**). Sample II, also a hypoechoic nodule with regular margins, measures about 18 x 10 mm (**a**). Like Sample I, it shows peripheral and central vascularity on both color Doppler (**b**) and Angio pl.u.s (**c**), and provides elastography parameters (**d**) for evaluation.

where

- $Z_{1-\alpha/2}$ is the critical value from the standard normal distribution corresponding to the desired confidence level,
- σ is the standard deviation,
- d is the margin of error.

Study workflow

The proposed approach (Figure 3) comprises two primary pathways. The first pathway involves the computation of TIRADS and elastographic parameters by expert radiologists. The second pathway focuses on automated medical image processing and artificial intelligence, incorporating preprocessing, feature extraction, and classification. Both pathways are designed to evaluate the diagnostic performance of human experts and AI in thyroid nodule assessment.

Radiologist-based process

In the first pathway, both senior and junior radiologists assess the thyroid images and determine TIRADS scores. The radiologists compute TIRADS features such as composition, shape, margin, echogenicity, calcification pattern, vascularity pattern, and microvascularity. elastography was conducted by a senior radiologist, and the features, including Emax, Emean, Emin, and the Q-box ratio, were directly extracted from the images.

Following this, we used ROC univariate statistical analysis on the elastographic parameters to determine their efficacy in diagnosing malignant nodules. The performance of the radiologists' assessments is evaluated using area under the curve(AUC), sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (PLR), negative likelihood ratio (NLR), and overall accuracy.

AI-based process

The second pathway centers on an automated ML approach. This includes preprocessing, feature extraction, and classification using AI. The workflow begins with the following preprocessing steps:

- **Image normalization:** The B-mode thyroid ultrasound images undergo min-max normalization to ensure consistency across all data. This method scales the pixel intensity values to a range between 0 and 1, following the formula:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

where X is the original pixel value, $X_{\min}$ is the minimum pixel value in the image, and $X_{\max}$ is the maximum pixel value.
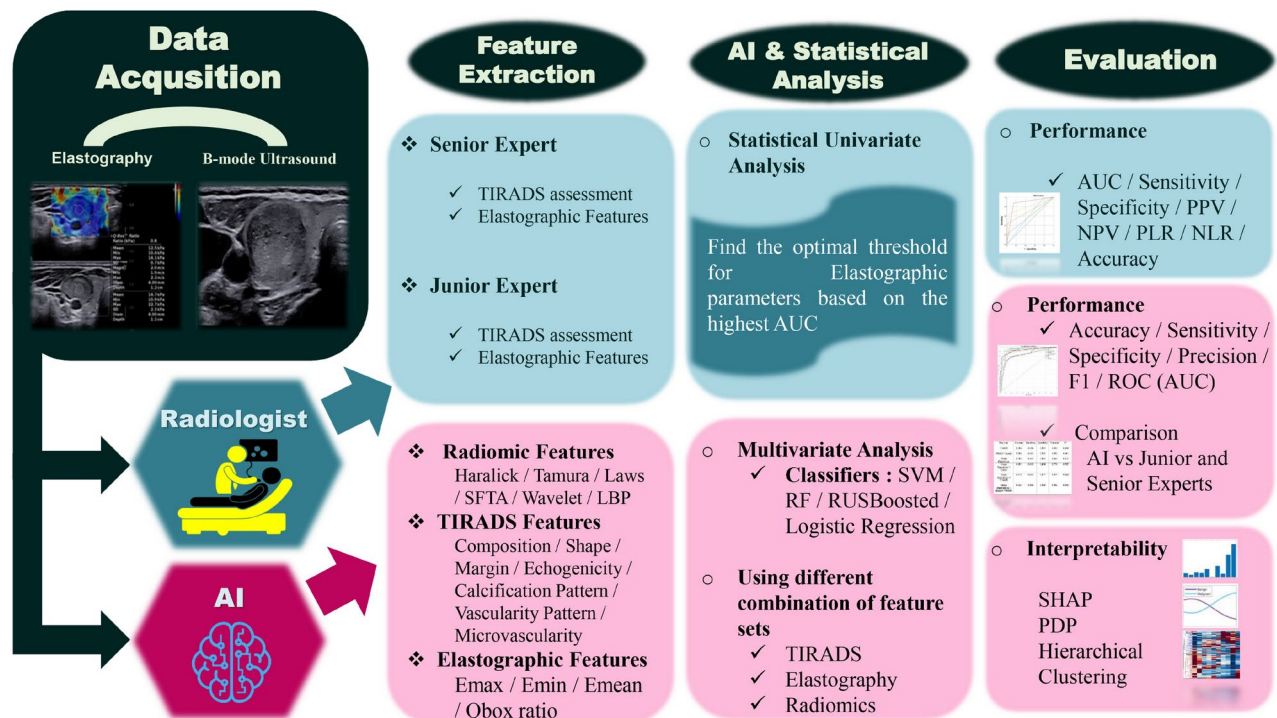
Fig. 3. It provides an informative visual representation of the procedural workflow employed in this research endeavor.
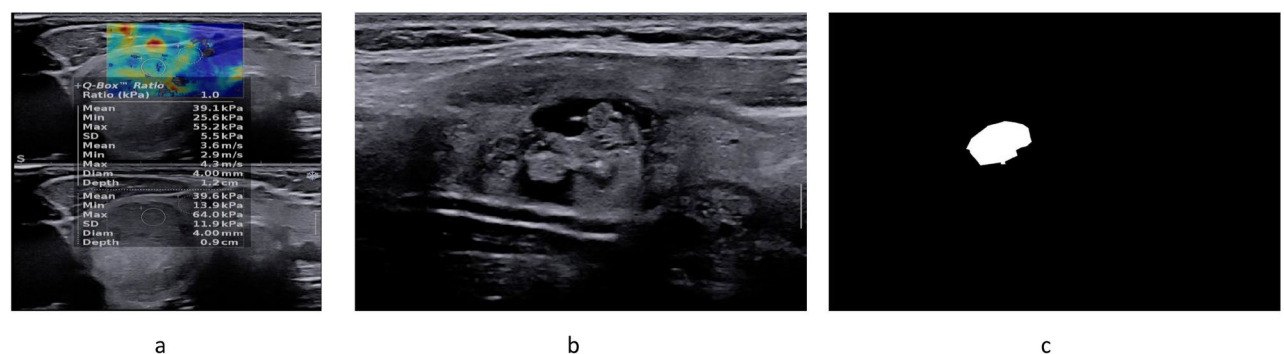

a   b   c

Fig. 4. It shows (**a**) elastography parameters, (**b**) normalized image and (**c**) segmented image.

- **Segmentation:** The Region of Interest (ROI), which focuses solely on the thyroid nodule, is manually selected by radiologists. By isolating the nodule, redundant information from surrounding tissues is removed, ensuring that the extracted features focus only on the nodule itself.

Figure 4 demonstrates the preprocessing and data preparation for the subsequent feature extraction.

After preprocessing, a wide range of Radiomic features is meticulously extracted from the segmented thyroid images[44]. These features provide a detailed analysis of texture and structure, helping to differentiate between benign and malignant nodules. Haralick, Tamura, Laws, Segmentation-based Fractal Texture Analysis (SFTA), Local Binary Pattern (LBP), and wavelet features are the radiomic features utilized in the study.

Classification & feature combinations
After feature extraction, several ML classifiers including: Support Vector Machines (SVM), Random Forest (RF), RUS Boosted trees, and Logistic Regression, are applied to evaluate the classification performance using different feature combinations. Each model was trained using an 80-20 train-test split, where 80% of the dataset was used for training and 20% for testing. Since each patient contributed only one nodule to our dataset, this partitioning introduced no data leakage. To ensure robust model performance, we applied 5-fold cross-validation, where the training set was divided into 5 equal parts, and the model was trained and validated across these partitions.

The seven combinations of the features are:

- Radiomic features only
- TIRADS features only
- Elastography features only
- Radiomic + TIRADS features
- Radiomic + Elastography features
- TIRADS + Elastography features
- Radiomic + TIRADS + Elastography features

Each combination undergoes classification, and performance is evaluated using accuracy, sensitivity, specificity, precision, F1 score, and AUC (ROC).

Analysis & interpretability

To assess feature quality, contribution, and validate the model's performance and reliability, we conducted various analyses aimed at ensuring both robustness and interpretability in classifying thyroid nodules.

SHAP analysis

To interpret the predictive contributions of TIRADS and elastography features, SHAP analysis was applied after model predictions were made[41]. SHAP values allowed us to understand the relative importance of each TIRADS and elastographic feature in the final classification of thyroid nodules as malignant or benign. This analysis particularly highlights whether elastography features provided additional predictive value beyond traditional ultrasonographic criteria.

The SHAP value for a feature j can be calculated using the following formula:

$$SHAP_i(j) = \sum_{S \subseteq N \setminus \{i\}} \left(\frac{|S|!(|N| - |S| - 1)!}{|N|!} \right) (f(S \cup \{i\}) - f(S))$$

where:

- $SHAP_i(j)$: SHAP value for feature j for instance i.
- $f(S)$: The model's predicted output for the set of features S.
- N: The complete set of features.
- S: A subset of features excluding feature i.
- $|S|$: The number of features in subset S.
- $|N|$: The total number of features.

PDP for elastography features

PDPs are used to visualize the relationship between specific elastographic features-such as Emax, Emean, Emin, and Q-box ratio[42]. PDPs demonstrate how the predicted probability for each class changes as the values of these features vary.

The underlying concept of PDPs can be mathematically represented by the formula:

$$PDP_j(x_j) = \frac{1}{N} \sum_{i=1}^{N} f(x_j, x_{-j}^{(i)})$$

where:

- $PDP_j(x_j)$: Represents the partial dependence of feature j at value x_j.
- f: The model's predicted output, indicating the likelihood of each class based on the feature values.
- N: The total number of instances in the dataset.
- $x_{-j}^{(i)}$: The feature values for all other features except for j for instance i.

Hierarchical clustering

Hierarchical clustering was employed to uncover relationships between features and samples[43]. This method provides a visual representation of patterns and correlations among TIRADS and elastographic features in our dataset. The clustering process begins by calculating the Euclidean distance between features and selecting the two closest features to group together. This is followed by progressively grouping additional features based on proximity, ultimately creating larger clusters until all features are categorized into distinct groups. The final output is a hierarchical structure that reveals the relationships between features and samples.

By applying this approach, we can effectively cluster our samples into two distinct groups based on their feature profiles. Analyzing the labels for each sample allows us to identify which samples fell into which group. If certain samples are mis-clustered, it suggests that their attributes do not conform to the general characteristics of their respective classes, indicating potential exceptions that warrant further investigation.

This analysis provides a quick overview of how well our features perform in distinguishing between classes. Furthermore, identifying the number of mis-clustered samples can offer insights into the robustness of our predictions, guiding us to focus on cases that may exhibit unusual behavior.

Permutation testing for model validation

To validate the performance of the ML model and ensure the reliability of the results, a permutation test is conducted 2000 times on the test dataset. This test is performed in two stages. In the first stage, the labels are randomly shuffled, and the classification is repeated 1000 times to establish the expected chance level, which is around 65% due to class imbalance. In the second stage, the model is tested 1000 times using the correct labels to assess its actual performance.

Importantly, in both stages, the test data are shuffled for each iteration. This ensures that the test is performed on different data splits each time, thereby guaranteeing that the model's results are not dependent on any specific subset of the dataset. To validate the model's effectiveness and reliability, we expect the classification performance to consistently exceed the chance level established by the random label shuffling.

DeLong's test

DeLong's test is a statistical method used to compare the areas under the ROC curves of two or more models. It calculates whether the differences in AUCs between models are statistically significant. The test computes a p-value to determine if one model significantly outperforms another in terms of discriminative ability.

In this study, DeLong's test was used to assess the performance of SVM models built with different feature combinations. The resulting p-values indicate whether the differences in the AUCs of the models are significant, with a threshold of $p < 0.05$ considered to denote statistical significance.

Results

In total, 181 patients were enrolled in this study, with a mean age of 47.9 ± 12.2 years (range: 16–82 years). The majority of the patients were female (86.2%), and 35.9% of the nodules were classified as malignant.

The nodules had a mean larger diameter of 18.1 ± 11.1 mm (range: 5–58 mm) and a smaller diameter of 12.8 ± 7.4 mm (range: 2–40 mm). Elastography measurements showed an average Emax of 49.4 ± 47 kPa (range: 5–300 kPa), Emin of 16.5 ± 13.5 kPa (range: 3.2–92 kPa), and Emean of 30.2 ± 30 kPa (range: 5–208 kPa). The mean Qbox ratio was 1.47 ± 1.38 (range: 0.3–8.8).

In terms of nodule location, 86 nodules (47.5%) were right-sided, 79 (43.6%) were left-sided, 15 (8.3%) were isthmic, and 1 (0.6%) was ectopic (lingual). Regarding the nodules' ultrasound features, the most common composition was "solid or predominantly solid" (85.1%), and over half of the nodules were hyper- or iso-echoic (59.1%). The majority of nodules had a wider-than-tall shape (87.8%), smooth margins (84%), no calcification (73.5%), and peripheral vascularity (61.9%). Among the malignant nodules, 97% were diagnosed as papillary thyroid carcinoma (PTC).

We compared the mean age, nodule size, Emax, Emin, Emean, and Qbox ratio between benign and malignant nodules. P-values indicate the likelihood that the observed differences occurred by chance, with smaller values confirming greater statistical significance. Table 2 shows the summary of these comparisons, with significant differences noted in all measured variables. Malignant nodules showed higher elasticity values across all categories, particularly for Emax, with a mean of 76.12 kPa compared to 34.13 kPa for benign nodules (P<0.001).

Additionally, we analyzed the frequency of malignancy in both genders, finding rates of 36.5% (57/156) in females and 32% (8/25) in males (P=0.82). This indicates no significant gender difference in malignancy rates.

Next, we compared the distribution of imaging descriptors between benign and malignant nodules, as shown in Table 3. The results indicated that specific features, such as solid or predominantly solid composition, hypoechoic echogenicity, and irregular margins, were associated with significantly higher malignancy rates. For instance, hypoechoic nodules presented a malignancy rate of 75.4%, while hyper- or iso-echoic nodules had a rate of only 7.5% (P<0.001).

	Pathology	Mean	Standard Deviation	P-Value
Age (year)	Benign	49.68	10.74	0.003
	Malignant	44.64	13.94	
Size1 (mm)	Benign	20.45	11.46	<0.001
	Malignant	13.83	8.97	
Size2 (mm)	Benign	14.26	7.65	<0.001
	Malignant	10.28	6.19	
Emax (kPa)	Benign	34.13	21.76	<0.001
	Malignant	76.12	64.60	
Emean (kPa)	Benign	20.83	13.13	<0.001
	Malignant	46.70	42.07	
Emin (kPa)	Benign	13.28	8.68	<0.001
	Malignant	22.26	17.83	
Qbox ratio	Benign	0.97	0.62	<0.001
	Malignant	2.34	1.84	

Table 2. Age, nodule size, Emax, Emean, Emin, and Qbox ratio between the benign and malignant nodules.

Sonographic Descriptor	Feature	Malignancy Percentage [No. of Total]	P-Value
Composition	Solid or Predominantly Solid	41.6 [64/154]	<0.001
	Mixed Solid and Cystic	5.6 [1/18]	
	Spongiform	0 [0/9]	
Echogenicity	Hyper- or Iso-echoic	7.5 [8/107]	<0.001
	Hypoechoic	75.4 [46/61]	
	Very Hypoechoic	84.6 [11/13]	
Shape	Wider than Tall	30.2 [48/159]	<0.001
	Round	50 [3/6]	
	Taller than Wide	87.5 [14/16]	
Margin	Smooth	24.3 [37/152]	<0.001
	Lobulate/Irregular	96.6 [28/29]	
CalcificationPattern	None	17.3 [23/133]	<0.001
	Macrocalcification	40 [2/5]	
	Peripheral/Rim	50 [1/2]	
	Punctate	94.4 [34/36]	
	Punctate and Macrocalcification	100 [4/4]	
	Punctate and Coarse	100 [1/1]	
VascularityPattern	None	68.8 [11/16]	0.022
	Central	20 [1/5]	
	Peripheral	33.9 [38/112]	
	Both	33.3 [15/45]	
Microvascularity	None	66.7 [8/12]	0.07
	Central	25 [1/4]	
	Peripheral	37 [34/92]	
	Both	32.7 [18/55]	
Junior TIRADS	2	4 [1/25]	<0.001
	3	7.8 [4/51]	
	4	24.2 [8/33]	
	5	72.2 [52/72]	
Senior TIRADS	2	0 [0/17]	<0.001
	3	2.7 [2/74]	
	4	37.9 [11/29]	
	5	85.2 [52/61]	

Table 3. Distribution of malignancy in different features of sonographic descriptors.

Test Result Variable(s)	AUC	P-Value	Asymptotic 95% CI	
			Lower Bound	Upper Bound
Emax	0.765	<0.001	0.692	0.838
Emean	0.760	<0.001	0.686	0.834
Emin	0.668	<0.001	0.580	0.756
Qbox ratio	0.801	<0.001	0.733	0.869

Table 4. Univariate evaluation of Emax, Emean, Emin, and Qbox ratio.

Statistical analysis

In the univariate analysis, we evaluated the diagnostic performance of elastographic features in distinguishing malignant from benign nodules. ROC analysis was conducted to calculate the AUC for each variable, which reflects its diagnostic efficacy. Table 4 presents these AUC values, along with p-values and 95% confidence intervals for Emax, Emean, Emin, and the Qbox ratio.

We also determined the optimal cut-off points for the elastographic features using the Youden index. Among the evaluated features, the Q-box ratio demonstrated the best performance at its selected cut-off point of 1.2, with a sensitivity of 75.4%, specificity of 73.7%, PPV of 62.0%, NPV of 84.0%, PLR of 2.86, NLR of 2.99, and overall accuracy of 74.3%. This highlights the Q-box ratio as the most diagnostically effective elastographic feature.

For sonographic descriptors (categorical variables), we grouped the variables into two subgroups. One subgroup contained features associated with higher malignancy rates, and the other subgroup contained

Variable	AUC	P-Value	Asymptotic 95% CI	
			Lower Bound	Upper Bound
Composition a	0.604	0.020	0.523	0.686
Echogenicity b	0.865	<0.001	0.806	0.925
Shape c	0.599	0.027	0.509	0.689
Margin	0.711	<0.001	0.625	0.797
Calcification d	0.797	<0.001	0.721	0.873

Table 5. Diagnostic indices of sonographic descriptors. a: Composition, Spongiform or mixed solid and cystic as benign vs. Solid or predominantly solid as malignant; b: Echogenicity, Hyper- or Iso-echoic nodules as benign vs. hypo-echo or very hypoechoic as malignant; c: Shape, Round or wider than tall as benign vs. taller than wide as malignant; d: Calcification, No calcification vs. any type of calcification.

Classifier	Accuracy	Sensitivity	Specificity	Precision	F1
Logistic Regression	0.861	0.769	0.876	0.833	0.800
Random Forest	0.867	0.784	0.883	0.836	0.809
SVM (RBF)	**0.922**	**0.892**	**0.939**	**0.892**	**0.892**
RUS Boosted Trees	0.861	0.861	0.917	0.777	0.817

Table 6. Performance evaluation metrics for different classifiers. Highest values are in bold.

Features	Accuracy	Sensitivity	Specificity	Precision	F1	AUC
Elasto	0.784	0.707	0.834	0.696	0.702	0.800
TIRADS	0.867	0.784	0.883	0.836	0.809	0.900
TIRADS+Elasto	0.889	0.815	0.900	0.869	0.841	0.930
Radiomics	0.785	0.815	0.881	0.662	0.731	0.835
Radiomics+Elasto	0.851	0.815	0.894	0.779	0.797	0.880
Radiomics +TIRADS	0.917	0.846	0.917	**0.917**	0.880	0.930
Radiomics+Elasto +TIRADS	**0.923**	**0.892**	**0.940**	0.892	**0.892**	**0.940**

Table 7. Performance evaluation metrics for SVM with different feature combinations. Highest values are in bold.

features with lower malignancy rates. For instance, the composition descriptor, with spongiform or mixed solid and cystic nodules (associated with 0% and 5.6% malignancy rates), was categorized as benign, while solid or predominantly solid nodules (41.6% malignancy rate) were classified as malignant.

After this categorization, the diagnostic performance of these sonographic descriptors was evaluated for malignancy prediction, with results shown in Table 5.

Among the evaluated sonographic descriptors, echogenicity demonstrated the best overall diagnostic performance. It achieved a sensitivity of 87.7%, specificity of 85.3%, PPV of 77.0%, NPV of 92.5%, PLR of 5.98, NLR of 6.93, and overall accuracy of 86.2%.

Notably, composition showed the highest sensitivity at 98.5%, with specificity of 22.4%, PPV of 41.6%, NPV of 96.3%, PLR of 1.26, NLR of 14.56, and overall accuracy of 49.7%.

Model evaluation

Table 6 summarizes the performance evaluation metrics, including accuracy, sensitivity, specificity, precision, and F1 score for four classifiers: logistic regression, random forest, SVM (RBF), and RUS Boosted trees. Among these classifiers, SVM (RBF) exhibited superior performance across all metrics. Performance metrics reported are based on the evaluation of the test dataset.

In addition, Table 7 provides a comprehensive evaluation of performance metrics for various feature combinations using SVM, including accuracy, sensitivity, specificity, precision, and F1 score. The combination of **Radiomics + TIRADS + Elastography** achieves the highest accuracy (0.923) and overall best performance across all metrics.

Furthermore, we performed DeLong's test to compare the performances of the SVM models built with different feature combinations. The p-values for the comparison of the SVM models with different feature combinations are presented in Table 8. These results suggest that the models incorporating combinations of different features show significant differences in performance.

Figure 5 evaluates the performance of different combinations of elastography and other features using ROC curves to distinguish between AI results, senior experts, and junior experts.

Feature Combinations	Elasto	TIRADS	TIRADS + Elasto	Image Features	Image Features + TIRADS	Image Features + Elasto
Elasto	-	0.4947	0.0752	0.2449	0.0909	0.8606
TIRADS	0.4947	-	0.0475	0.0799	0.1690	0.5985
TIRADS + Elasto	0.0752	0.0475	-	0.0042	0.9808	0.0904
Image Features	0.2449	0.0799	0.0042	-	0.0053	0.0817
Image Features + TIRADS	0.0909	0.1690	0.9808	0.0053	-	0.1277
Image Features + Elasto	0.8606	0.5985	0.0904	0.0817	0.1277	-
Image Features + TIRADS + Elasto	0.0438	0.0820	0.6510	0.0023	0.2031	0.0727

Table 8. P-values for pairwise comparisons of SVM models built with different feature combinations are presented. A significance threshold of 0.05 was used to determine statistical significance.
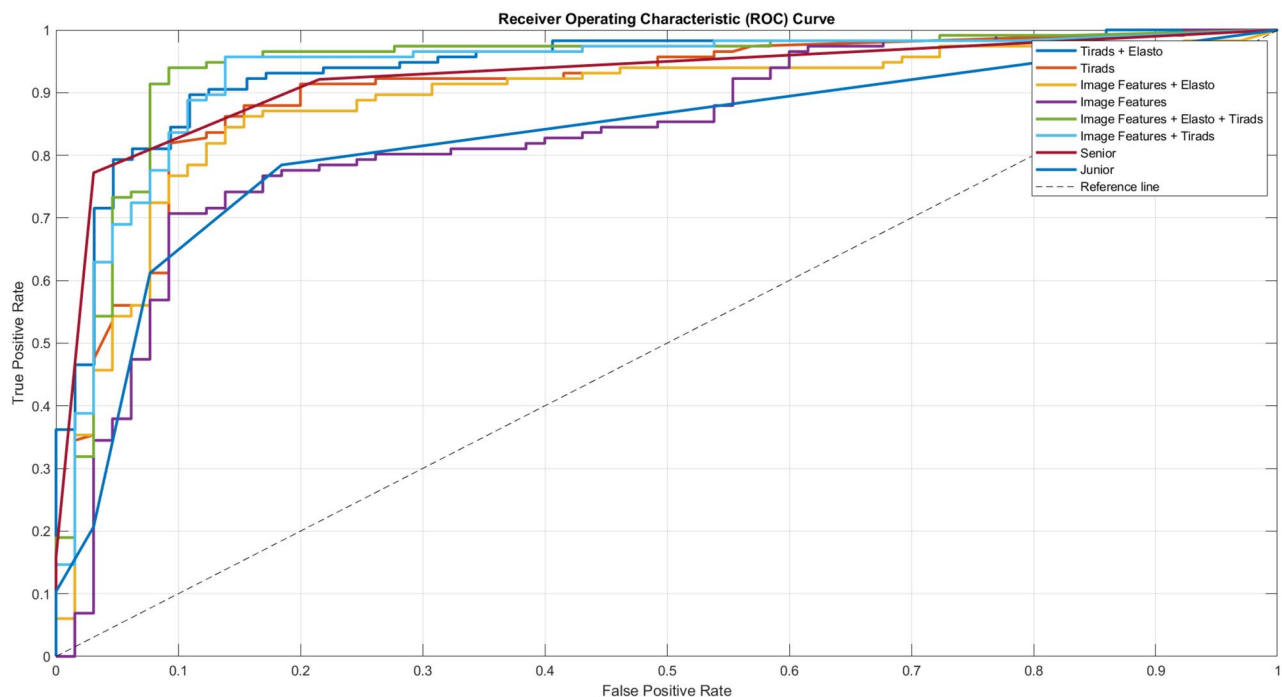
Fig. 5. ROC curves comparing AI, senior experts, and junior experts in performance assessment.

Figure 6 presents the results of the permutation test conducted to assess the model's accuracy. The histogram distributions indicate that chance-level accuracy is centered around 0.62, while the model's accuracy is centered around 0.91. Statistical analysis yielded a non-significant p-value (< 0.0001), suggesting the model's accuracy significantly exceeds the chance level.

Interpretable ML analysis

In this section, we provide a comprehensive analysis of feature importance and classification performance using three key techniques: SHAP values, PDP, and hierarchical clustering. These methods collectively help us understand the contributions of various features to the classification of benign and malignant nodules, while also enhancing physicians' understanding of the AI's decision-making process by improving interpretability.

Fig. 7 illustrates the SHAP values for elastographic and TIRADS features, highlighting the significant role of elastographic features, particularly the Q-box ratio, in the model's predictions. Each score displayed in the figure represents the average SHAP value for that feature across all samples in each class. For the malignant class on the right, the score is the average of the SHAP scores calculated for the 65 malignant samples. Similarly, for the benign class on the left, the score reflects the average of the SHAP scores computed for the 116 benign samples. This method of aggregation provides insight into how each feature contributes to the classification of nodules as benign or malignant.

The PDPs further reveal the relationship between elastographic feature values and malignancy likelihood. As shown in Figure 8, an increase in elastographic feature values correlates with a rising likelihood of malignancy, while lower values are associated with the benign class. This relationship underscores the diagnostic importance of elastographic measurements in identifying malignant lesions.
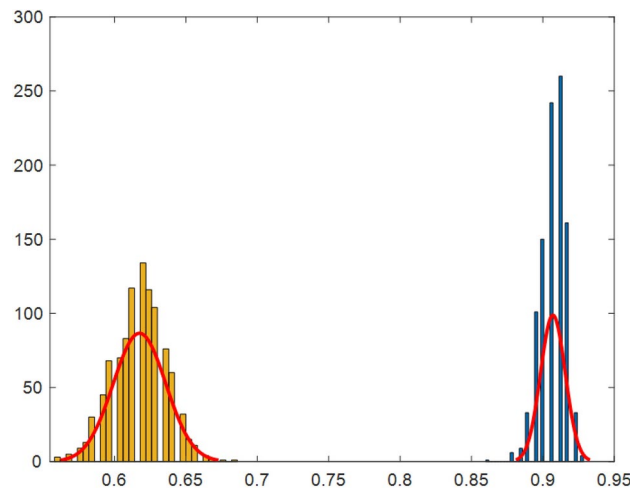
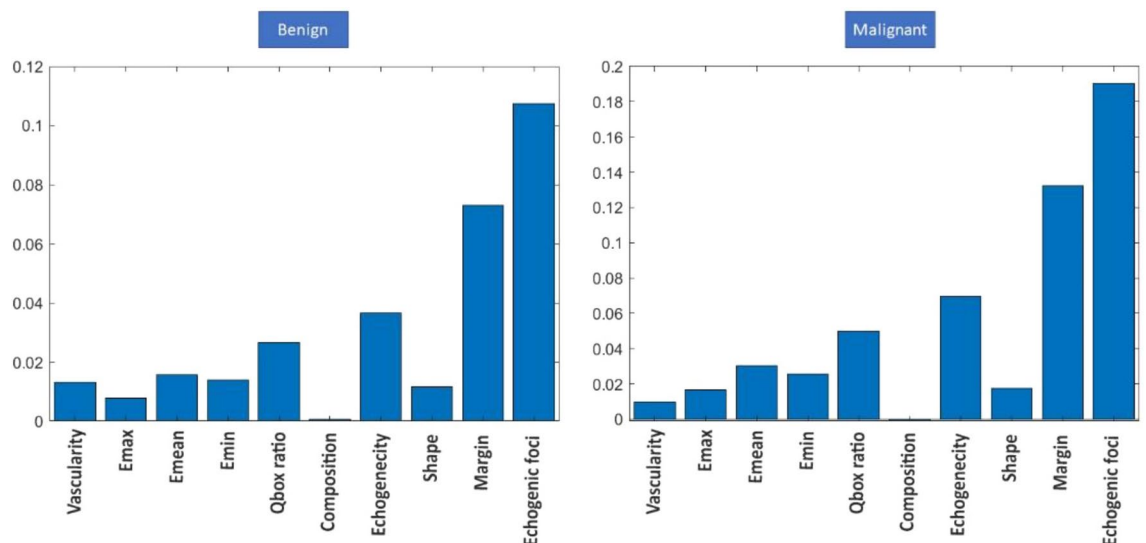**Fig. 6**. Results of the permutation test to evaluate model accuracy.

Fig. 7. SHAP values for elastographic and TIRADS features.

In Figure 9, the hierarchical clustering results are presented, visually distinguishing two clear groups of samples. The red box corresponds to the first cluster, while the blue box represents the second. The top dendrogram depicts the relationships among features, and the left dendrogram showcases the similarities among samples. This clustering method illustrates how the samples form two distinct groups based solely on the elastographic and TIRADS features, reinforcing their relevance in malignancy classification.

This unsupervised analysis reflects the human decision-making process used by radiologists, who also rely on these features to differentiate between benign and malignant nodules. The black and yellow boxes indicate misclustered samples, which are particularly noteworthy as they do not conform to the typical patterns of their respective groups. These exceptions serve as warnings for practitioners, indicating instances where the standard features could lead to inaccurate predictions. Understanding these anomalies can offer valuable insights into potential factors that influence misdiagnosis.

Discussion

This study aimed to enhance the accuracy of thyroid nodule malignancy prediction by integrating TIRADS, elastography, and ultrasound-based radiomics. While ML methods have previously been employed for thyroid nodule detection, incorporating elastography into this diagnostic framework is a novel approach. Our method combines multiple diagnostic modalities, offering a comprehensive assessment that can assist physicians in distinguishing between malignant and benign nodules.

Several prior studies have explored the application of radiomics and ML in predicting thyroid nodules. However, only a few have incorporated elastography as a parameter. Shayesteh et al.[23] evaluated sonographic
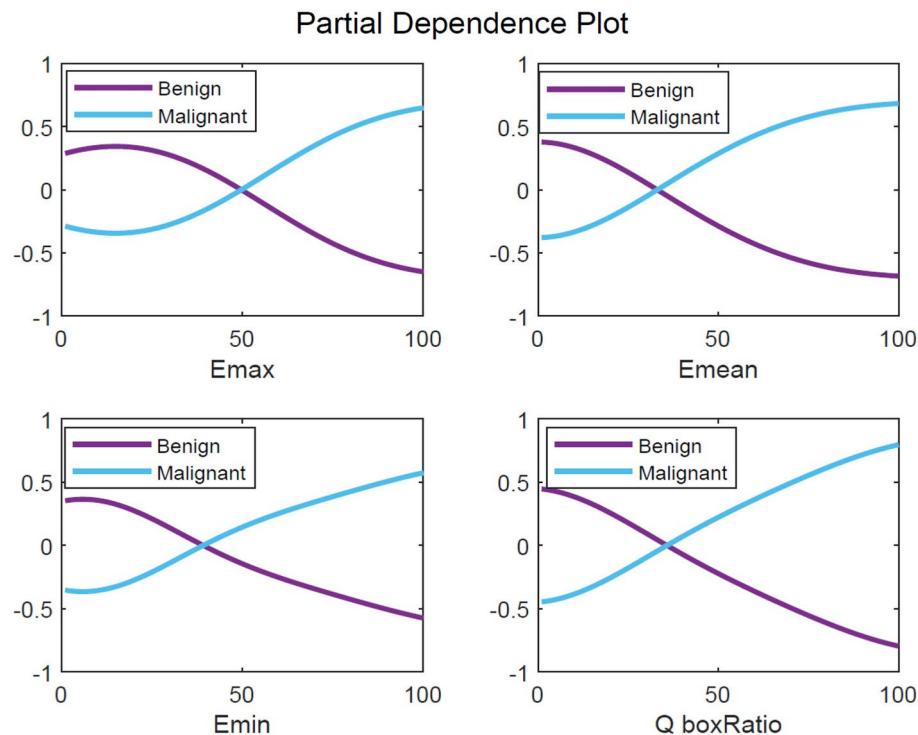
Fig. 8. PDP for elastographic features and malignancy likelihood.
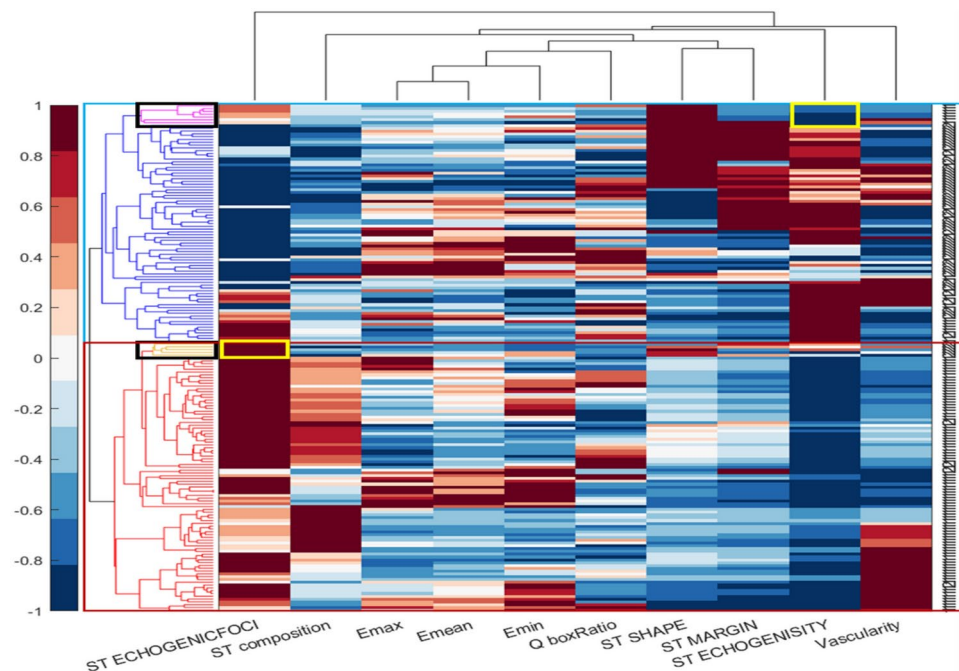

Fig. 9. Hierarchical diagram showing feature relationships and sample similarities.

image radiomic features as biomarkers for thyroid nodule malignancy in patients recommended for FNA, achieving an accuracy of 0.70, sensitivity of 0.81, and specificity of 0.58. Arabi et al.[24] investigated the use of radiomics in ultrasound images, demonstrating a higher accuracy of 0.89, sensitivity of 0.92, and specificity of 0.85 for malignancy prediction. Liang et al.[38] developed a radiomics score based on ultrasound imaging, yielding an AUC of 0.93 in the validation cohort, outperforming assessments made by less experienced radiologists. This

suggests that radiomics scores provide a robust predictive ability for thyroid nodule malignancy compared to evaluations by junior radiologists using the American College of Radiology (ACR) TIRADS system.

Colakoglu et al.[39] explored the ability of ML algorithms to differentiate between benign and malignant nodules using radiomic features, applying the random forest algorithm and achieving sensitivity of 85.2, specificity of 87.9, and accuracy of 86.8. Zhang et al.[40] employed four distinct ML models, including those based solely on ultrasound features and those incorporating both elastography and ultrasound features. Their results revealed diagnostic accuracy rates of 84.6, 85.7, 83.3, and 84.1, with sensitivities of 88.1, 89.1, 83.5, and 85.0, and specificities of 84.2, 85.3, 83.2, and 83.6, respectively. These findings underscore the potential of ML models to provide more reliable discrimination between benign and malignant nodules compared to radiologist diagnoses relying solely on ultrasound or on a combination of ultrasound and elastography.

Our use of SHAP values and PDPs revealed important insights into the ML model's decision-making process. Specifically, echogenic foci, margin, and echogenicity emerged as the most influential features in malignancy prediction, which aligns with clinical experience. Elastographic feature values were particularly correlated with malignancy likelihood, further validating their significance in diagnosis.

We performed DeLong's test to compare the performances of SVM models built with different feature combinations. The results, shown in Table 8, suggest that combining TIRADS with elastography significantly improves the model's performance compared to using TIRADS or Image features alone. Similarly, adding TIRADS (or TIRADS + Elastography) to Image features also yields a significant performance boost. However, when models already combine multiple features, such as TIRADS + Elasto or Image features + TIRADS + Elasto, the difference in performance compared to similar multi-feature models is not statistically significant. These findings imply that augmenting single-feature approaches (either TIRADS or Image features alone) with additional information (Elastography or a combination of TIRADS and Elastography) can lead to better diagnostic performance in this setting.

Hierarchical clustering provided an unsupervised analysis of how TIRADS and elastography features align with clinical outcomes. By identifying misclustered samples (highlighted in black and yellow in Figure 5), we highlighted cases where traditional diagnostic methods may face limitations. These misclustered cases, such as those shown in Figures 8 and 9, offer insights into discrepancies between AI predictions and expert evaluations. This provides opportunities for identifying outliers and refining diagnostic strategies.

In the specific case of Sample I in Figure 8, both senior and junior radiologists assigned it a TIRADS score of 5, indicating high suspicion for malignancy. However, AI analysis predicted the nodule as benign, a finding corroborated by histopathologic results. This discrepancy underscores the potential of AI to identify subtleties that may be missed by human evaluators, warranting further investigation into the diagnostic methods used in such cases.

In contrast, Sample II in Figure 8, identified as a spindle cell and medullary thyroid carcinoma (MTC), was accurately classified as malignant by both AI and expert radiologists, with the histopathologic diagnosis confirming this finding. The agreement across diagnostic methods, including elastographic parameters, highlights the value of combining these modalities in diagnosing complex cases like MTC, which can be challenging to identify based solely on traditional criteria.

Sample I in Figure 9 also revealed a malignant pattern in elastography, raising intermediate suspicion among radiologists. However, AI predicted the nodule to be benign, a result that conflicted with the final FNA report, which confirmed Papillary Thyroid Carcinoma (PTC). This divergence highlights the complexity of thyroid nodule diagnostics, especially when dealing with nodules exhibiting ambiguous features. It emphasizes the importance of incorporating both elastography and AI in the diagnostic process to reduce potential oversights.

In Sample II in Figure 9, all diagnostic modalities, including AI, radiologist evaluations, and elastography, indicated a benign nodule. This consensus was confirmed by histopathology, underscoring the reliability of integrating AI with traditional diagnostic tools to improve diagnostic confidence in cases where the nodule presents benign features.

Looking forward, further research should explore the integration of additional imaging modalities, such as photoacoustic imaging, in conjunction with ultrasound for thyroid nodule assessment. Investigations into combining elastographic features with radiomics for regression analysis may also yield valuable insights. Moreover, identifying specific biomarkers from this diverse set of features and analyzing correlations between TIRADS, elastography, and radiomics could enhance our understanding of thyroid nodule malignancy prediction.

Despite the promising results, our study has limitations. The relatively small sample size (N=181), along with data collection from a single site using a single scanner, may limit the generalizability of our findings. Future research should aim to validate these results across larger, more diverse datasets and assess the impact of combining elastographic features with other imaging techniques, such as photoacoustic imaging.

Overall, this study demonstrates that integrating TIRADS, elastography, and radiomics can significantly enhance the prediction of thyroid nodule malignancy. These techniques alongisde ML models present an opportunity for reducing diagnostic discrepancies and optimizing patient care, especially in challenging cases where human evaluation alone may fall short.

Conclusion

The objective of this study was to enhance the detection of malignant thyroid nodules by integrating TIRADS, elastography, and advanced ML algorithms based on radiomics, culminating in the development of the ELTIRADS model. The findings highlight the substantial role of elastography in predicting malignancy. When combined with B-mode ultrasound and TIRADS criteria, elastography significantly improves diagnostic accuracy. Furthermore, the study proposes establishing an optimal threshold for elastography factors to refine the detection of thyroid nodule malignancy. Additionally, it introduces an interpretable approach using hierarchical

clustering, SHAP, and PDP. This approach assists physicians in distinguishing benign from malignant nodules and potentially enhancing clinical diagnostic practices.

Data availibility

Code availability

References
1. WM T: The spectrum of thyroid disease in a community: the Whickham survey. Clin Endocrinol (Oxf) 7:481-493, 1977
2. Vander, J. B., Gaston, E. A. & Dawber, T. R. The significance of nontoxic thyroid nodules: final report of a 15-year study of the incidence of thyroid malignancy. *Annals of internal medicine* **69**, 537–540 (1968).
3. Shi, M., Nong, D., Xin, M. & Lin, L. Accuracy of Ultrasound Diagnosis of Benign and Malignant Thyroid Nodules: A Systematic Review and Meta-Analysis. *International Journal of Clinical Practice* **2022**, 5056082 (2022).
4. Durante, C. et al. The natural history of benign thyroid nodules. *Jama* **313**, 926–935 (2015).
5. Durante, C. et al. The Diagnosis and Management of Thyroid Nodules: A Review. *Jama* **319**, 914–924 (2018).
6. Stoian D, Cornianu M, Dobrescu A, Lazăr F: Nodular thyroid cancer. Diagnostic value of real time elastography. Chirurgia (Bucur) 107:39-46, 2012
7. Monpeyssen, H., Tramalloni, J., Poirée, S., Hélénon, O. & Correas, J. M. Elastography of the thyroid. *Diagn Interv Imaging* **94**, 535–544 (2013).
8. Rago, T., Santini, F., Scutari, M., Pinchera, A. & Vitti, P. Elastography: new developments in ultrasound for predicting malignancy in thyroid nodules. *J Clin Endocrinol Metab* **92**, 2917–2922 (2007).
9. Kwak, J. Y. et al. Thyroid imaging reporting and data system for US features of nodules: a step in establishing better stratification of cancer risk. *Radiology* **260**, 892–899 (2011).
10. Tessler, F. N. et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. *J Am Coll Radiol* **14**, 587–595 (2017).
11. Nikiforova, M. N. & Nikiforov, Y. E. Molecular diagnostics and predictors in thyroid cancer. *Thyroid* **19**, 1351–1361 (2009).
12. Kumar, V. et al. Automated Segmentation of Thyroid Nodule, Gland, and Cystic Components From Ultrasound Images Using Deep Learning. *IEEE Access* **8**, 63482–63496 (2020).
13. Sabouri, M. et al. Machine learning based readmission and mortality prediction in heart failure patients. *Scientific Reports* **13**, 18671 (2023).
14. Daskareh, M. et al. Predicting Rheumatoid Arthritis Development Using Hand Ultrasound and Machine Learning-A Two-Year Follow-Up Cohort Study. *Diagnostics* **14**, 1181 (2024).
15. Taleie H, et al.: Left Ventricular Myocardial Dysfunction Evaluation in Thalassemia Patients Using Echocardiographic Radiomic Features and Machine Learning Algorithms. Journal of digital imaging:1-13, 2023
16. Khodabakhshi, Z. et al. Overall survival prediction in renal cell carcinoma patients using computed tomography radiomic and clinical information. *Journal of digital imaging* **34**, 1086–1098 (2021).
17. Mohebi M, et al.: Post-revascularization Ejection Fraction Prediction for Patients Undergoing Percutaneous Coronary Intervention Based on Myocardial Perfusion SPECT Imaging Radiomics: a Preliminary Machine Learning Study. Journal of Digital Imaging:1-16, 2023
18. Sabouri, M. et al. Myocardial Perfusion SPECT Imaging Radiomic Features and Machine Learning Algorithms for Cardiac Contractile Pattern Recognition. *Journal of Digital Imaging* **36**, 497–509 (2023).
19. Amini, M. et al. Overall survival prognostic modelling of non-small cell lung cancer patients using positron emission tomography/computed tomography harmonised radiomics features: The quest for the optimal machine learning algorithm. *Clinical Oncology* **34**, 114–127 (2022).
20. Shiri I, et al.: Diagnosis of COVID-19 using CT image radiomics features: a comprehensive machine learning study involving 26,307 patients. medRxiv:2021.2012. 2007.21267367, 2021
21. Khodabakhshi, Z. et al. Non-small cell lung carcinoma histopathological subtype phenotyping using high-dimensional multinomial multiclass CT radiomics signature. *Computers in biology and medicine* **136**, 104752 (2021).
22. Gharibi O, et al.: Maximizing the Predictive Power of Radiomic Features in SPECT Images: A Comparative Study of Reconstruction Algorithms Using Machine Learning. Proc. 2023 IEEE Nuclear Science Symposium, Medical Imaging Conference and International Symposium on Room-Temperature Semiconductor Detectors (NSS MIC RTSD): City
23. Shayesteh SP, et al.: Machine Learning Based Malignancy Prediction in Thyroid Nodules Malignancy: Radiomics Analysis of Ultrasound Images. Proc. 2020 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC): City
24. arabi M, et al.: A machine learning-based sonomics for prediction of thyroid nodule malignancies. Endocrine 82:326-334, 2023
25. Xi, N. M., Wang, L. & Yang, C. Improving the diagnosis of thyroid cancer by machine learning and clinical data. *Scientific Reports* **12**, 11143 (2022).
26. Alyas, T. et al. Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach. *BioMed Research International* **2022**, 9809932 (2022).
27. Gomes Ataide, E. J. et al. Thyroid Nodule Classification for Physician Decision Support Using Machine Learning-Evaluated Geometric and Morphological Features. *Sensors* **20**, 6110 (2020).
28. Zhang, B. et al. Machine Learning-Assisted System for Thyroid Nodule Diagnosis. *Thyroid* **29**, 858–867 (2019).
29. Wang, L. et al. Multimodal Ultrasound Radiomic Technology for Diagnosing Benign and Malignant Thyroid Nodules of Ti-Rads 4–5: A Multicenter Study. *Sensors* **24**, 6203 (2024).
30. Chen, J. H. et al. Applying machine-learning models to differentiate benign and malignant thyroid nodules classified as C-TIRADS 4 based on 2D-ultrasound combined with five contrast-enhanced ultrasound key frames. *Front Endocrinol* **15**, 1299686 (2024).
31. Ioachim, D. The Bethesda System for Reporting Thyroid Cytopathology. *Acta Endocrinol (Buchar)* **14**, 282–283 (2018).
32. Moon, H. J. et al. Diagnostic performance of gray-scale US and elastography in solid thyroid nodules. *Radiology* **262**, 1002–1013 (2012).
33. Haralick RM, Shanmugam K, Dinstein IH: Textural features for image classification. IEEE Transactions on systems, man, and cybernetics:610-621, 1973
34. Dash, S. & Jena, U. R. Multi-resolution Laws' Masks based texture classification. *Journal of applied research and technology* **15**, 571–582 (2017).

35. Pitas I: Digital image processing algorithms and applications: John Wiley & Sons, 2000
36. Costa AF, Humpire-Mamani G, Traina AJM: An efficient algorithm for fractal analysis of textures. Proc. 2012 25th SIBGRAPI Conference on Graphics, Patterns and Images: City
37. Prasad L, Iyengar SS: Wavelet analysis with applications to image processing: CRC press, 1997
38. Liang, J. et al. Predicting malignancy in thyroid nodules: radiomics score versus 2017 American College of Radiology Thyroid Imaging. *Reporting and Data System. Thyroid* **28**, 1024–1033 (2018).
39. Colakoglu B, Alis D, Yergin M: Diagnostic value of machine learning-based quantitative texture analysis in differentiating benign and malignant thyroid nodules. Journal of oncology 2019, 2019
40. Zhang, B. et al. Machine learning-assisted system for thyroid nodule diagnosis. *Thyroid* **29**, 858–867 (2019).
41. Lundberg SM, Lee S-I: A unified approach to interpreting model predictions. Advances in neural information processing systems 30, 2017
42. Friedman JH: Greedy function approximation: a gradient boosting machine. Annals of statistics:1189-1232, 2001
43. Sørlie, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences* **98**, 10869–10874 (2001).
44. Mayerhoefer, M. E. et al. Introduction to Radiomics. *J Nucl Med* **61**, 488–495. https://doi.org/10.2967/jnumed.118.222893 (2020).

Acknowledgements

Declarations

Competing interests
The authros declare no competing interests.

Ethical approval
This prospective study was approved by the Research Council of Tehran University of Medical Sciences (Ethical code, IR.TUMS.AMIRALAM.REC.1399.042).

Consent to participate
Informed consent was waived by ethic groups.

Consent to publish
Informed consent was waived by ethic groups.

Additional information
Correspondence and requests for materials should be addressed to M.D., H.S. or M.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.