European Journal of Radiology Open 16 (2026) 100716

Contents lists available at ScienceDirect

European Journal of Radiology Open

journal homepage: www.elsevier.com/locate/ejro

Machine Learning for diagnosis of malignant thyroid nodules based on thyroid ultrasound: Systematic review and meta-analysis of studies with external datasets

Elisa Gatta [a,b], Roberto Gatta [b,c], Riccardo Morandi [b,d], Samuele Isoli [a], Sara Corvaglia [e], Simone Vetrugno [e], Virginia Maltese [a], Ilenia Pirola [a,b], Claudio Casella [b,d], Carlo Cappelli [a,b,*]

[a] Department of Clinical and Experimental Sciences, SSD Endocrinologia, University of Brescia, ASST Spedali Civili, Brescia, Italy
[b] Centro per la Diagnosi e Cura delle Neoplasie Endocrine e delle Malattie della Tiroide, University of Brescia, Brescia, Italy
[c] Department of Clinical and Experimental Sciences, University of Brescia, Brescia, Italy
[d] Department of Clinical and Experimental Sciences, Surgical Clinic, University of Brescia, ASST Spedali Civili, Brescia, Italy
[e] Department of Internal Medicine and Therapeutics, University of Pavia, Pavia, Italy

HIGHLIGHTS

- AI shows high accuracy in thyroid nodule malignancy detection.
- Pooled sensitivity 87 % and specificity 83 % across 27 studies.
- Cytology- and histology-based validations yielded comparable diagnostic performance.
- Results support AI integration into thyroid nodule management.

ARTICLE INFO

Keywords:
Thyroid nodules
Machine learning
Artificial intelligence

ABSTRACT

Introduction: Optimizing the diagnostic approach to thyroid nodules remains a crucial challenge. Ultrasound-based risk stratification systems such as EU-TIRADS have shown reasonable sensitivity and specificity. Therefore, we conducted a systematic review and meta-analysis to assess the diagnostic performance of Artificial Intelligence (AI) models in differentiating benign from malignant thyroid nodules on ultrasound data.

Methods: A comprehensive search of PubMed/MEDLINE, Scopus, and Web of Science was performed up to January 1, 2025. Eligible studies included patients with thyroid nodules undergoing ultrasound, where AI-based models were validated against cytological or histological findings. The AI algorithms were developed using different types of ultrasound-derived data, including B-mode images, radiomics features. Pooled sensitivity, specificity, and area under the curve (AUC) were estimated using a hierarchical summary receiver operating characteristic (HSROC) model.

Results: Twenty-seven studies comprising 146,332 patients and over 600,000 ultrasound images met inclusion criteria. Overall, pooled sensitivity was 87 % (95 % CI: 84–89 %) and specificity 83 % (95 % CI: 79–86 %). The summary operating point indicated a sensitivity of 88 % and specificity of 83 %, with an AUC of 91.9 % (95 % CI: 90.0–93.2 %). Although subgroup analysis suggested higher accuracy when cytology was used as the reference standard compared to histology, the mixed-effects meta-regression did not confirm a statistically significant association ($p = 0.238$ for sensitivity; $p = 0.188$ for specificity).

Conclusion: AI-based algorithms show excellent diagnostic performance in distinguishing benign from malignant thyroid nodules, with robust validation across external datasets. These findings support the potential integration of AI into clinical thyroid nodule management, although further multicenter, non-Asian, and histology-based studies are warrantee.

Systematic review registration: PROSPERO (CRD420251108149)

* Correspondence to: Department of Clinical and Experimental Sciences, SSD Endocrinologia, Universiy of Brescia, ASST Spedali Civili Brescia, Piazzale Spedali Civili n°1, Brescia 25121, Italy.
E-mail address: carlo.cappelli@unibs.it (C. Cappelli).

https://doi.org/10.1016/j.ejro.2025.100716
Received 24 August 2025; Received in revised form 27 November 2025; Accepted 3 December 2025

1. Introduction

Ultrasound is the cornerstone imaging modality for the evaluation of thyroid owing to its wide availability, safety, and real-time capability [1,2]. Conventional B-mode imaging provides high-resolution morphological information, allowing assessment of echogenicity, margins, and internal architecture. Doppler imaging complements this by evaluating vascular patterns, which may reflect tissue perfusion and inflammatory or neoplastic activity [3]. More recently, elastography techniques—both strain and shear-wave—have enabled quantitative assessment of tissue stiffness, offering additional diagnostic insights [4–6]. Thyroid nodules (TNs) are highly prevalent in the general population, particularly among women and older adults [7]. Given their frequency and the potential for underlying malignancy, international guidelines uniformly advocate prompt ultrasound risk stratification at diagnosis [8]. Despite these strengths, ultrasound remains inherently operator-dependent and subject to artifacts that may limit reproducibility and inter-observer agreement [9].

Moreover, conventional ultrasound-based risk stratification systems, although standardized, still rely on subjective visual interpretation and may yield inconsistent performance across centers and operators. The inter-observer variability in describing sonographic features such as echogenicity or margins remains a key limitation of the current approach to thyroid nodule management [9].

The optimization of diagnostic and therapeutic strategies has long been a central objective in medical practice. In the field of thyroidology, particularly in the selection of nodules for cytological evaluation, several ultrasound-based scoring systems have been developed to identify lesions suitable for fine-needle aspiration, demonstrating satisfactory sensitivity and specificity [8,10,11]. Among these, a recent meta-analysis by Yang et al. encompassing 88 studies and 59,304 nodules reported a sensitivity of 75 % and a specificity of 82 % for EU-TIRADS category TR5. Notably, specificity decreased substantially for categories TR4 and TR3, reaching 62 % and 31 %, respectively [12].

The concept of using computers to simulate intelligent behavior, later termed AI, was first introduced by Alan Turing in the 1950s [13]. Since then, this field has undergone a dramatic evolution, with profound implications for medicine. In particular, the application of AI to medical imaging has been proposed as a strategy to enhance diagnostic accuracy, consistency, and efficiency. In 2017, the U.S. Food and Drug Administration approved the first cloud-based deep learning application for clinical use in healthcare [14], marking a turning point in the integration of AI into routine practice. Deep learning has since been applied to lesion detection, differential diagnosis generation, and automated reporting, impacting almost all areas of medicine [14]. In thyroidology, one of the earliest applications of AI to ultrasound imaging was reported in the 1990s, when Karakitsos et al. employed a back-propagation neural network to assist clinicians in distinguishing between benign and malignant thyroid nodules [15].

AI is increasingly applied in medicine, offering promising opportunities for improving diagnostic accuracy and workflow efficiency, although its clinical adoption remains limited. Its applications extend well beyond professional domains, influencing everyday life and clinical practice alike. AI was first applied to thyroidology in the 1990s, and its potential is now being increasingly investigated across a growing body of studies [16,17].

Radiomics, a promising branch of AI, enables the automated extraction of large volumes of quantitative data from medical images. In the context of thyroid nodules, radiomics has demonstrated the ability to detect imaging features imperceptible to the human eye, thereby improving risk stratification, supporting differentiation between benign and malignant lesions, and informing decision-making regarding fine-needle aspiration. By integrating imaging-derived data with advanced algorithms and predictive models, radiomics may promote a more accurate, personalized, and non-invasive approach to thyroid nodule management [18–20]. However, current AI-based models also present relevant challenges, including the need for large annotated datasets, lack of model interpretability (“black-box” nature), and limited generalizability across populations and imaging devices. Conversely, their main strength lies in the ability to quantify subtle imaging features beyond human perception, potentially supporting more objective and reproducible risk stratification (BIBLIO).

The aim of this systematic review and meta-analysis was to quantitatively assess the diagnostic accuracy of AI models—validated on external, independent datasets—in differentiating benign from malignant thyroid nodules, using cytological or histological diagnosis as the reference standard.

2. Material and methods

2.1. Search strategy and inclusion criteria

A wide literature search of the PubMed/MEDLINE, Scopus and Web of Science databases was made.

The review questions were defined according to the diagnostic accuracy framework (Population, Index test, Reference standard, and Diagnosis):

- Population: patients with thyroid nodules who underwent ultrasound examination;
- Index test: artificial intelligence–based models (including machine learning and radiomics) applied to ultrasound data;
- Reference standard: cytological and/or histological findings;
- Diagnosis (target condition): discrimination between benign and malignant thyroid nodules.

The algorithm used for the research was the following: (“artificial intelligence” OR “machine learning” OR “deep learning” OR “neural network” OR “decision tree” OR “random forest” OR “nearest neighbor” OR “naive Bayes” OR “support vector machine”) AND (“diagnosis” OR “screen*” OR “classifi*” OR “discriminat*” OR “performance” OR “sensitivity” OR “specificity” OR “accuracy” OR “area under the curve” OR “AUC” OR “calibrat*”) AND (“thyroid” OR “thyroid gland”) AND (“cancer” OR “neoplasm*” OR “carcinoma” OR “nodule*” OR “tumor*” OR “tumour*” OR “malignan*” OR “adenoma”).

The search was updated until January 1, 2025. Only articles in English were considered, and preclinical studies, conference proceedings, reviews, or editorials were excluded. To expand our search, the references of the retrieved articles were also screened for additional papers; moreover, other studies were identified by looking through all the articles that cite the papers included in the review ("snowballing"). All eligible studies were exported and managed in EndNote 20.3.

2.2. Eligibility criteria

The eligibility criteria were chosen taking into account the review question. Studies addressing the review questions defined according to the PICO framework were included. Exclusion criteria for the systematic review (qualitative analysis) were reviews, letters, comments, or editorials on the topic of interest, on the analyzed topic (as these articles are characterized by poor-quality evidence and are typically affected by publication bias), and original studies unrelated to the review question, such as those evaluating AI models applied to cytology, pathology, or non-ultrasound imaging data.

Studies were included if they assessed the diagnostic performance of artificial intelligence models using ultrasound-based inputs with cytological or histological confirmation as the reference standard. All indeterminate or equivocal diagnoses, whether histological or cytological (e.g., Bethesda III and IV), were excluded from the analysis. In addition, studies in which the diagnostic reference standard consisted solely of an ultrasound-based risk score (i.e., TI-RADS) without cytological and/or histological confirmation were excluded.

To ensure methodological homogeneity, studies in which clinical data were integrated as a major component of the model were excluded.

2.3. Study selection

E.G. and C.C. independently read the titles and abstracts of the records generated by the search algorithm. They then determined which studies were eligible based on predefined criteria. Discrepancies were resolved through discussion, and a third reviewer (R.G.) was consulted in case of disagreement.

E.G. is currently pursuing a Ph.D. in Artificial Intelligence and obtained her medical degree five years ago; she has also recently completed her specialization in Endocrinology. C.C. is a board-certified endocrinologist with twenty years of specialty experience and currently serves as Professor of Endocrinology. R.G. holds a Ph.D. in Oncological Science and serves as a researcher, with primary research interests in radiomics and process mining for healthcare.

2.4. Reporting

The protocol of this systematic review is registered in PROSPERO (CRD420251108149) and followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [21] (Fig. 1). The quality assessment of the studies, including the risk of bias and applicability concerns, was carried out using Quality Assessment of Diagnostic Accuracy Studies version 2 (QUADAS-2) evaluation [22] (Fig. 2). The methodological quality of the included studies was evaluated using the METhodological RadiomICs Score (METRICS) [23] (Fig. 3).

2.5. Data extraction

The reviewers collected data from all included studies by examining the full text, tables, and supplementary materials. Extracted data covered general study information (authors, publication year, country, study design, clinical setting), patient characteristics (sample size, inclusion/exclusion criteria, relevant clinical features), reference standard (histological and/or cytological confirmation), type of internal validation, number of images, and automated nodule localization. The main findings of the articles included in this review are reported in the Results section

2.6. Statistical analysis

To assess the diagnostic performance of the machine learning algorithm, pooled meta-analyses of sensitivity and specificity were conducted using a bivariate random-effects model. In addition, pooled diagnostic odds ratio (DOR), positive likelihood ratio (DLR+), and negative likelihood ratio (DLR–) were calculated. A mixed-effects meta-regression model was applied to explore potential moderators accounting for between-study heterogeneity, specifically to assess differences in diagnostic performance between studies adopting cytological versus histological outcomes as the reference standard. Publication bias was assessed using Deeks' funnel plot asymmetry and the associated weighted regression test. A hierarchical summary receiver operating characteristic (HSROC) curve was generated, including pooled sensitivity, specificity, area under the curve (AUC) values with corresponding 95 % confidence intervals (CIs), summary operating point and prediction regions. All analyses were performed using RStudio (version

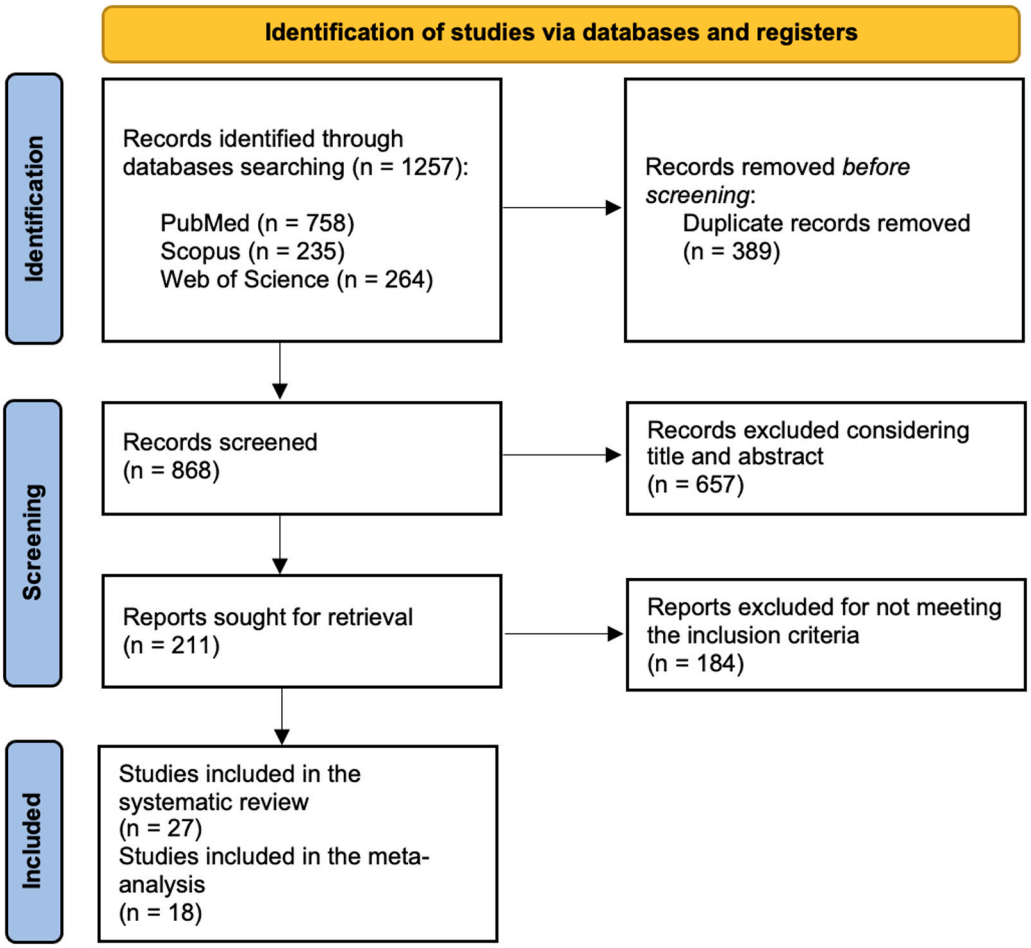

Fig. 1. Flowchart of the study selection process for eligible studies.

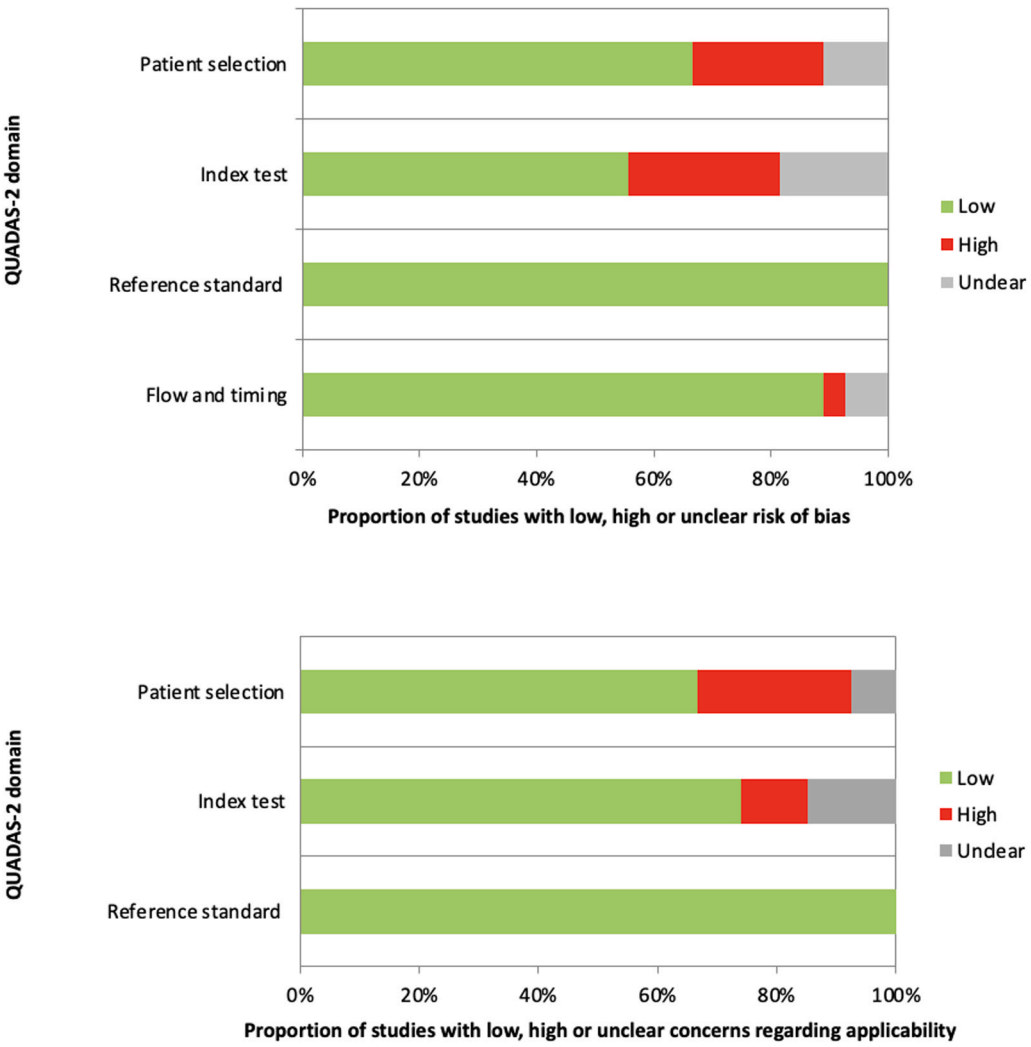

Fig. 2. QUADAS-2 assessment of risk of bias and applicability concerns for the studies included in the review.

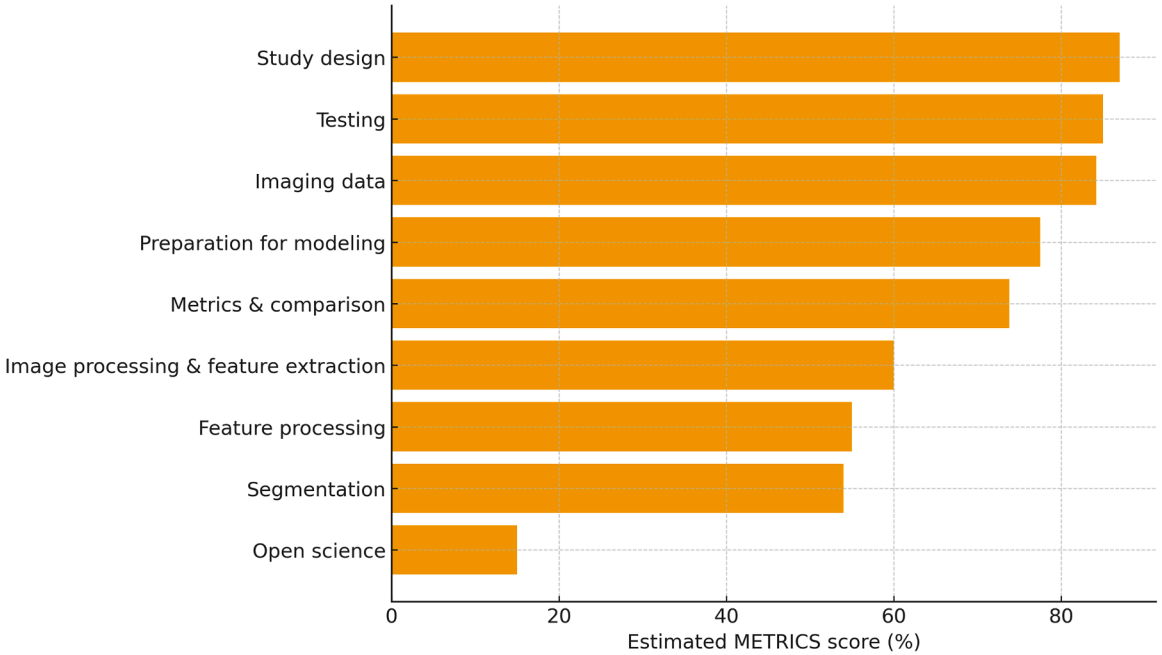

Fig. 3. Distribution of METhodological RadiomICs Score (METRICS) domain scores across the included studies.

2025.05.1 +513), with a significance level set at $p < 0.05$.

3. Results

3.1. Literature search

A total of 1257 articles were identified through the computerized literature search. After removal of duplicates, 868 articles remained. By screening titles and abstracts, 657 articles were excluded as their content was not relevant to the focus of this review. "Consequently, 211 articles were retrieved for full-text evaluation, of which 184 were excluded because they did not meet the inclusion criteria of our study. In particular, most of them lacked an external dataset for algorithm validation (Fig. 1). Therefore, 27 studies were ultimately included in the review [24–50]. Among these, only 18 studies provided complete diagnostic data (true positives, false negatives, false positives, and true negatives), allowing the construction of 2 × 2 contingency tables and their inclusion in the quantitative meta-analysis.

In general, the quality assessment using QUADAS-2 evaluation underlined the presence of unclear risk of bias and applicability concerns in some of the studies for what concerns patients' selection, index test, reference standard and flow and timing (Fig. 2). Several studies did not clearly report whether patients were enrolled consecutively or randomly; in addition, exclusion criteria were often poorly described, contributing to the high risk of bias in this domain. Moreover, several papers did not specify whether the threshold for malignancy was prespecified or derived post-hoc, which represents a methodological concern in diagnostic accuracy studies. The quality assessment using the METRICS revealed moderate overall methodological rigor among the included studies, with the highest scores observed in the domains of study design and testing, and lower adherence in the areas of segmentation, feature processing, and open-science practices (Fig. 3).

3.2. Characteristics of the studies

The main characteristics and results of the included studies are summarized in Table 1. In total, 146,332 patients were evaluated. Of the 27 studies, 24 (133,813 patients) were retrospective, 2 (11,755 patients) were retrospective–prospective and 1 prospective (764 patients). Eighteen studies compared machine learning performance against histopathological findings, while nine used cytological outcomes from fine-needle aspiration as the reference standard.

In all eligible studies, sonographic data were retrospectively extracted from institutional picture archiving and communication systems (PACS). The vast majority of AI models were trained on B-mode ultrasound images representing the largest longitudinal and/or transverse planes of each thyroid nodule. When applicable, the region of interest (ROI) was either manually or automatically segmented before feature extraction or network training. Approximately one third of the studies implemented radiomics pipelines to derive quantitative descriptors of texture, shape, and gray-level distribution from these ROIs.

Overall, 524,438 images were used for training, and 100,084 images for external validation. All studies were conducted in Asia.

When studies were stratified by imaging modality and AI framework, distinct performance patterns emerged, revealing how algorithm design and ultrasound input influence diagnostic reliability.

Models trained on static B-mode images using conventional convolutional neural networks (CNNs) achieved pooled AUCs ranging from 0.83 to 0.94, with sensitivity typically between 85–93 % and specificity between 82–94 % [29,36]. These models demonstrated robust discrimination, especially in large multicenter datasets, yet their outputs remained largely black-box and prone to variability when applied across ultrasound systems or acquisition protocols.

By contrast, knowledge-guided architectures integrating TI-RADS features [24,32] or interpretable frameworks [46] maintained comparable accuracy (AUC 0.90–0.93) but introduced an important conceptual advance: they linked each AI-derived probability to interpretable sonographic descriptors such as composition, echogenicity, or margins. This mechanistic transparency mitigates interobserver heterogeneity and enhances trust in AI-assisted decision support, representing a crucial step toward regulatory acceptance.

Three studies directly compared AI-assisted ultrasound with conventional sonographic interpretation by radiologists using TI-RADS or equivalent descriptors. In Chen et al., the AI model for thyroid nodule classification based on multitask deep learning using TI-RADS characteristics achieved an AUC of 0.91, sensitivity of 83 %, and specificity of 87 %, which was comparable to experienced radiologists (AUC 0.93; sensitivity 92 %) but with significantly higher specificity (80 %, $p = 0.02$), and clearly superior to junior readers (AUC 0.78; sensitivity 70 %; specificity 75 %) [34]. Similarly, in an effective and result-interpretable and-to-and thyroid nodule classification network, the AI system outperformed both intermediate and senior radiologists on internal and external validation datasets, with statistically significant differences in AUC according to DeLong's test [36]. Another study employing an ensemble deep-learning classifier (EDLC-TN) reported AI accuracy and AUC equal to or higher than those of individual radiologists, with further improvement when AI outputs were integrated with human interpretation [29].

Studies using multimodal or video-based ultrasound further improved performance by incorporating dynamic and contextual cues—such as vascularity, tissue elasticity, and temporal consistency—achieving the highest pooled accuracy (AUC ≈ 0.97, sensitivity ≈ 0.90, specificity ≈ 0.94) [49]. These results indicate that temporal information, often neglected in static-image datasets, carries incremental diagnostic value by approximating real-time expert evaluation.

Finally, transformer-based models leveraging self-attention mechanisms showed stable generalization across centers, sexes, and ultrasound devices (AUC ≈ 0.94) [48]. Their performance consistency suggests that attention-based architectures may capture higher-order spatial dependencies less sensitive to vendor-related variation.

3.3. Quantitative analysis

Although 18 studies were included in the quantitative synthesis, the number of datapoints shown in the forest plots (Figs. 4 and 5) is higher. This discrepancy arises because several studies assessed the same AI model on multiple independent external datasets, each yielding separate diagnostic performance metrics. As reported in Table 1 (column 'I/E/E/E'), these datasets were extracted and analyzed individually to preserve their distinct outcomes. Accordingly, the total number of evaluations represented in the meta-analysis exceeds the number of included papers.

Overall, the aggregated sensitivity of all included studies was 87 % (95 % CI: 84–89 %) (Fig. 4), while the specificity was 83 % (95 % CI: 79–86 %) (Fig. 5). Based on these pooled estimates, the DLR+ was 5.1, the DLR– was 0.16, and the DOR was 32.7. The Deeks' funnel plot for publication bias (Fig. 6) shows a relatively symmetric distribution of studies around the regression line, suggesting the absence of major small-study effects. The regression test for funnel plot asymmetry confirmed this visual impression, yielding a non-significant result ($t = 0.67$; $p = 0.51$). The corresponding area under the curve (AUC) was 91.9 % (95 % CI: 90.0–93.2 %), with a summary operating point indicating a sensitivity of 88 % and a specificity of 83 % (Fig. 7).

To investigate potential bias related to the diagnostic reference standard (cytological vs. histological data), we performed a subgroup analysis.

Six studies (15,572 patients, 95,100 images) compared machine learning results with cytological outcomes [25,30,35,36,40,48,49]. The pooled sensitivity was 91 % (95 % CI: 88–93 %) (Fig. 8), and specificity was 87 % (95 % CI: 76–93 %) (Fig. 9). Based on these pooled estimates, the DLR+ was 7.0, the negative likelihood ratio DLR–was 0.10, and the DOR was 70.0. The Deeks' funnel plot (Fig. 10) illustrates a nearly

Table 1
Characteristics of the human studies considered for the review.

First author	Ref. N.	Year	Study design	N. of patients[a]	N. of images[a]	Reference standard (histological and/or cytological) for external testing set	AI Method used	Type of internal validation	Automated nodule localization
Li	[24]	2019	Retrospective study	44,070/ 154/1420	312,399/ 8606/NR/ 741; 11,039	Histological	DL – CNN	NA	Yes
Song	[25]	2019	Retrospective study	NA	1358/NR/ 55/100	Cytological	DL – TL	Holdout cross-validation	No
Song	[26]	2019	Retrospective study	1580/299	6228/NR/ 367/152	Histological	DL – CNN	Five-fold cross-validation	Yes
Bai	[27]	2020	Retrospective study	NA	14,531/NR/ 3633/437; 570	Histological	DL – RS-Net; CNN	Five-fold cross-validation	No
Koh	[28]	2020	Retrospective study	14,194/ 781/200/ 200	13,560/NR/ 634/781; 200; 200	Histological	DL – CNN	NA	No
Wei	[29]	2020	Retrospective study	11,604/261	17,859/NR/ 7650/1032	Histological	DL	Random split-sample validation	Yes
Zhou	[30]	2020	Retrospective study	1629/105	1097/547/ NR/105	Cytological	DL – CNN; TL	NA	Yes
Peng	[31]	2021	Retrospective–prospective study	8339/ 1428/ 1048/303	14,439/ 3610/NR/ 4305	Histological	DL – RS-Net	NA	Yes
Wu	[32]	2021	Retrospective study	1396/197	1146/NR/ 143/112 698/NR/95/ 101 1844/NR/ 238/213	Histological	DL – CNN	Random split-sample validation	No
Zhu	[33]	2021	Retrospective study	6426/261	16,401/ 1000/300/ 1032	Histological	DL – CNN	NA	Yes
Chen	[34]	2022	Retrospective study	450/186	1076/269/ NR/243	Histological	DL – CNN	Five-fold cross-validation	No
Deng	[35]	2022	Retrospective study	NA	3125/391/ 391/831	Histological and cytological	DL – CNN; MTL	Random split-sample validation	Yes
Han	[36]	2022	Retrospective study	3096/886	2344/781/ 781/886	Cytological	MTL DenseNet	Ten-fold cross-validation	No
Keutgen	[37]	2022	Retrospective study	NA	734/184/ 68/66	Histological and cytological	ML	Five-fold cross-validation	No
Kim	[38]	2022	Retrospective study	7518/59	12,327/ 3082/432/ 168	Histological	DL – CNN	Random split-sample validation	Yes
Zhang	[39]	2022	Retrospective study	NA/400/ 587	NR	Histological	DL – CNN	Ten-fold stratified cross-validation	No
Chen	[40]	2023	Retrospective study	485/80	1012/253/ NR/126	Histological and cytological	DL – CNN	Three-fold cross-validation	No
Gao	[41]	2023	Retrospective study	NA/NA	4989/NR/ NR/309	Histological	MTL	Five-fold cross-validation	No
Tang	[42]	2023	Retrospective study	NA	7700/1923/ NR/431	Histological	TS-DSANN	Five-fold cross-validation	No
Xu	[43]	2023	Retrospective study	10,023 (I+E)	18,477/ 4563/1904 (I+E)	Histological	AI	Random split-sample validation	No
Yang	[44]	2023	Retrospective study	432/71	1392/349/ NR/309	Histological	DL – CNN	Random split-sample validation	No
Yao	[45]	2023	Retrospective study	1349/163/ 178	1349/NR/ NR/163; 178	Histological	DL – CNN,	Ten-fold cross-validation	Yes
Yao	[46]	2023	Retrospective study	7460/619	8017/1002/ NR/1002	Histological	DL – CNN	NA	Yes
Chen	[47]	2024	Retrospective study	6401/138	7636/1097/ 2129/339	Histological	DL – CNN	NA	Yes
Feng	[48]	2024	Retrospective study	7881/574	16,906/NR/ 1130/1262	Histological and cytological	DL	NA	No
Wu	[49]	2024	Prospective study	672/92	38,336/ 5376/ 10,368/8064	Histological and cytological	DL – CNN, ST	NA	Yes
Zhou	[50]	2024	Retrospective–prospective study	346/291	NR/NR/977/ 906	Histological and cytological	AI	NA	No

Ref.: references; N.: number; AI: artificial intelligence; NA: not available; DL: deep learning; CNN: Convolutional neural network; TL: transfer learning; RS-Net: risk stratification network; MTL: Multitask Learning; TS-DSANN: texture and shape focused dual-stream attention neural network; ST: swin-transformer
[a] Internal test/external test/external test/external test

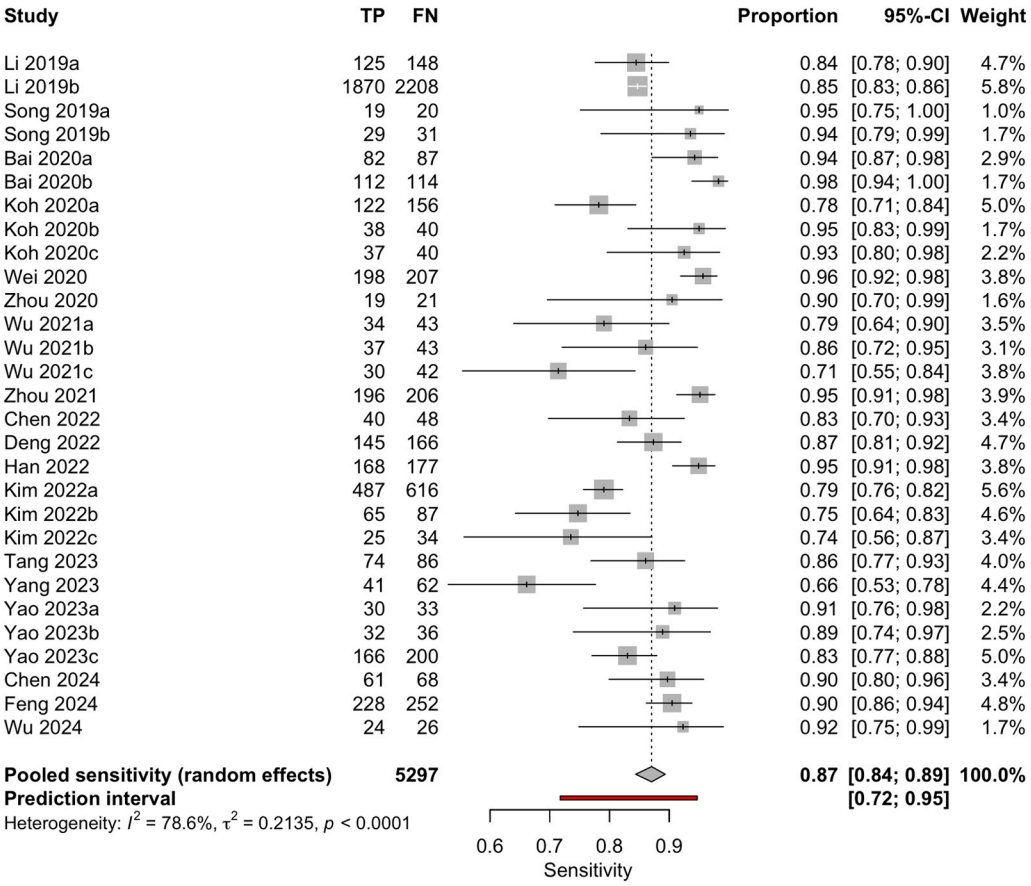

Fig. 4. Forest plot of sensitivity estimates from studies evaluating machine learning algorithms for the diagnosis of thyroid carcinoma, using either histological or cytological findings as the reference standard and tested on external datasets.

symmetric distribution of studies around the regression line, without substantial evidence of small-study effects. The regression test for funnel plot asymmetry confirmed the absence of significant publication bias (t = –1.26; p = 0.28). The AUC was 94.3 % (95 % CI: 92.4–96.0 %), with a summary operating point corresponding to a sensitivity of 92 % and a specificity of 87 % (Fig. 11).

Fourteen studies (108,067 patients, 474,383 images) compared results with histopathological findings [24,26–29,32–34,38,42,44–47]. The pooled sensitivity was 86 % (95 % CI: 83–88 %) % (Fig. 12), and specificity was 82 % (95 % CI: 77–86 %) (Fig. 13). Based on these pooled estimates, the DLR+ was 4.8, the DLR– was 0.17, and the DOR was 28.2. The Deeks' funnel plot (Fig. 14) shows a symmetric distribution of studies around the regression line, indicating the absence of relevant small-study effects or selective publication. The regression test for funnel plot asymmetry confirmed this visual impression, yielding a non-significant result (t = 0.07; p = 0.95). The AUC was 91.0 % (95 % CI: 89.7–93.2 %), with a summary operating point corresponding to a sensitivity of 87 % and a specificity of 82 % (Fig. 15).

Although the subgroup analysis suggested higher accuracy when cytology was used as the reference standard compared to histology, the mixed-effects meta-regression did not confirm a statistically significant association (p = 0.238 for sensitivity; p = 0.188 for specificity). The regression coefficient for logit sensitivity was β = –0.071 (95 % CI: –0.19–0.05) and for logit specificity β = –0.090 (95 % CI: –0.22–0.04). The residual heterogeneity was negligible for sensitivity ($\tau^2 = 0$) and low for specificity ($\tau^2 = 0.013$), indicating that the moderator variable did not meaningfully account for between-study variability.

3.4. Discussion

The present systematic review and meta-analysis highlights the excellent diagnostic performance of machine learning algorithms in differentiating benign from malignant thyroid nodules, with pooled sensitivities and specificities of 87 % and 83 %, respectively, validated in external datasets.

These findings align with the expanding literature supporting the application of AI in thyroid ultrasound imaging [51–54]. Jassal et al. recently highlighted the burgeoning potential of AI in the clinical management of cytologically indeterminate thyroid nodules; however, they emphasized that most available studies lacked robust, independent external validation [55]. In contrast, the present study, which focused on AI applications in differentiating benign from malignant thyroid nodules, demonstrated consistently good accuracy across external datasets.

To date, ultrasound-based risk stratification systems relying on nodular features have been established to identify lesions suspected of malignancy and to improve interobserver agreement. A large meta-analysis by Yang et al., including 88 studies and 59,304 nodules, reported a sensitivity of 75 % and a specificity of 82 % for EU-TIRADS category TR5. Notably, specificity declined substantially for TR4 and TR3 categories, reaching 62 % and 31 %, respectively [12]. In contrast, our study, which included all nodules regardless of their TIRADS

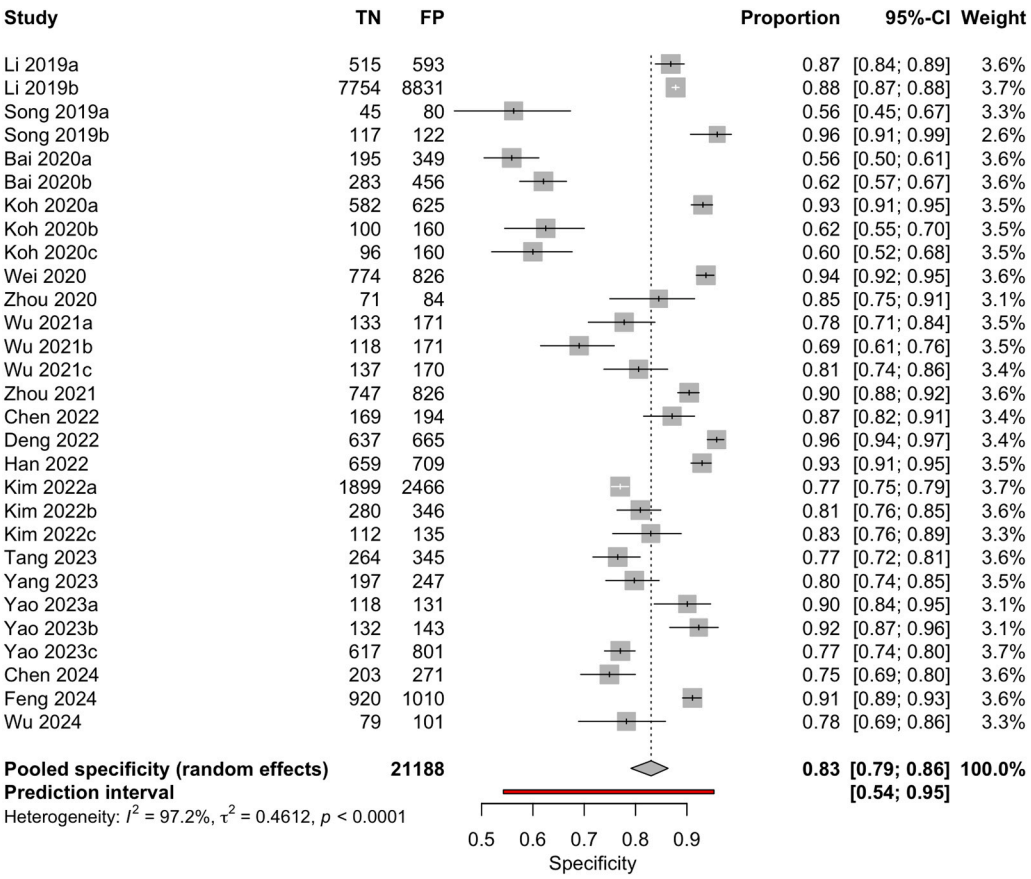

Fig. 5. Forest plot of specificity estimates from studies evaluating machine learning algorithms for the diagnosis of thyroid carcinoma, using either histological or cytological findings as the reference standard and tested on external datasets.

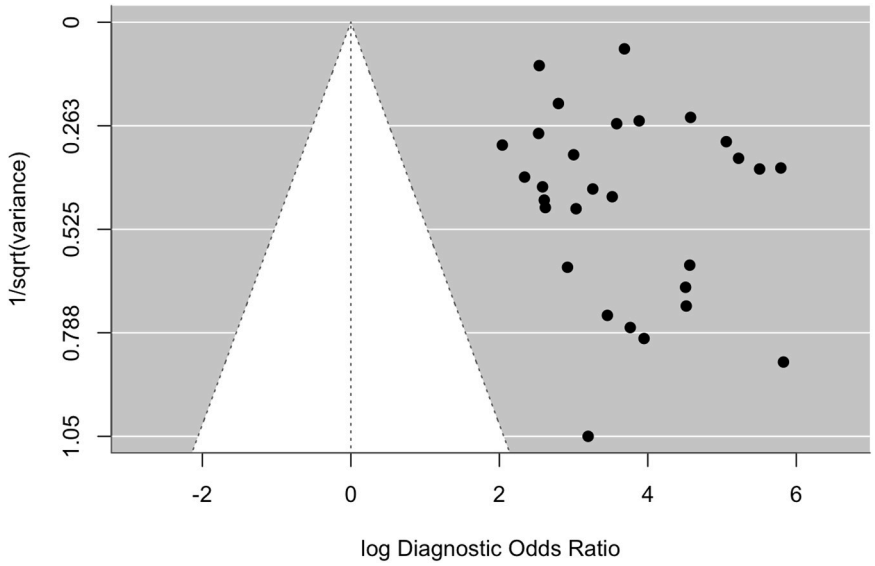

Fig. 6. Deeks' funnel plot for publication bias of studies using histological or cytological findings as the reference standard and tested on external datasets.

classification, demonstrated superior diagnostic performance. In addition, head-to-head evidence from included studies consistently showed AI matching or surpassing human readers in AUC, sensitivity, and specificity, highlighting its ability to reduce inter-observer variability and enhance diagnostic consistency across centers. The higher accuracy observed in our analysis likely reflects the ability of machine learning algorithms to capture complex textural and morphological features beyond human visual perception, thus improving reproducibility and reducing interobserver variability. Importantly, this approach is not subject to the selection bias that may arise when lesions are categorized solely on the basis of visual ultrasound assessment.

When comparing machine learning performance according to the reference standard adopted (cytological versus histological findings), no statistically significant differences were demonstrated. Although slightly

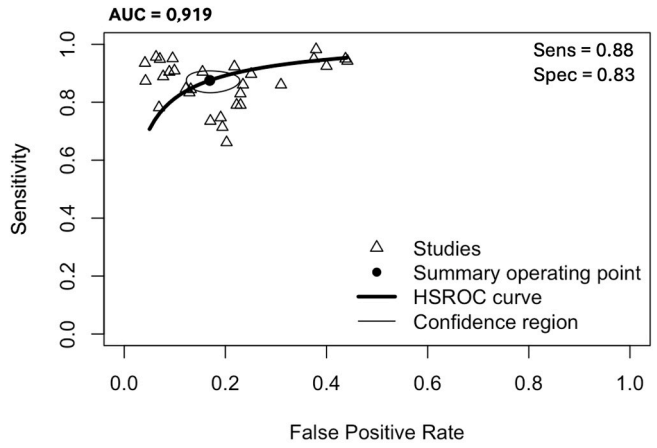

Fig. 7. Hierarchical summary receiver operating characteristic curve of machine learning algorithms for the diagnosis of thyroid carcinoma, using histological or cytological findings as the reference standard and tested on external datasets.

higher pooled estimates were observed in studies using cytology, the mixed-effects meta-regression analysis did not confirm a significant association with diagnostic accuracy (p = 0.238 for sensitivity; p = 0.188 for specificity). This result suggests that the choice of reference standard did not substantially influence the overall diagnostic performance of the evaluated models, supporting the robustness of machine learning algorithms across different validation settings.

From a clinical perspective, AI could improve the selection of nodules requiring FNA, thereby reducing unnecessary procedures, associated complications, and healthcare costs. Although FNAC is generally considered simple, reliable, safe, and well-accepted by patients, it is not entirely free from complications [56]. Minimizing its overuse remains an important goal in precision thyroid care.

Despite these promising results, several methodological and clinical issues limit large-scale implementation. Most existing algorithms are trained on single-center or ethnically homogeneous datasets, potentially reducing their applicability to broader populations and imaging devices. Furthermore, the intrinsic "black-box" nature of deep learning models continues to hinder interpretability and clinical trust, as the relative contribution of sonographic features—such as microcalcifications, vascularity, or stiffness—remains largely opaque.

Nevertheless, AI offers clear advantages over conventional risk stratification systems, including improved standardization, quantitative feature extraction, and enhanced diagnostic consistency. When adequately validated, these tools could assist radiologists in achieving

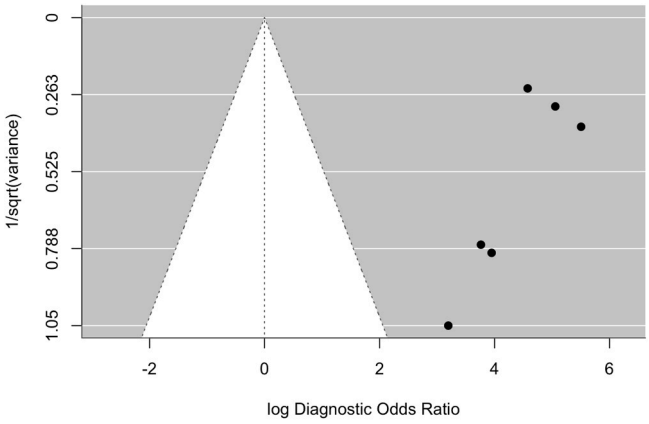

Fig. 10. Deeks' funnel plot for publication bias of studies using cytological findings as the reference standard and tested on external datasets.

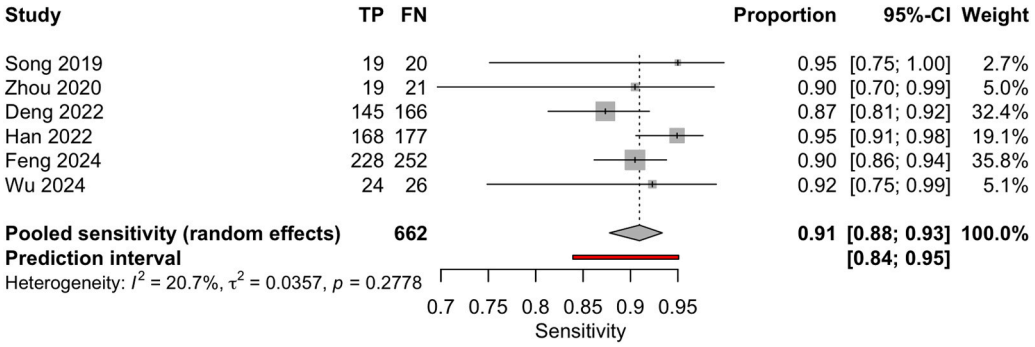

Fig. 8. Forest plot of sensitivity estimates from studies evaluating machine learning algorithms for the diagnosis of thyroid carcinoma, using cytological findings as the reference standard and tested on external datasets.

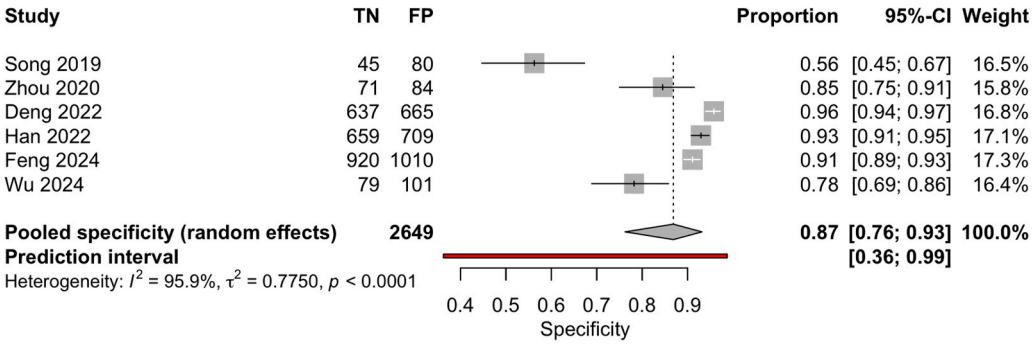

Fig. 9. Forest plot of specificity estimates from studies evaluating machine learning algorithms for the diagnosis of thyroid carcinoma, using cytological findings as the reference standard and tested on external datasets.

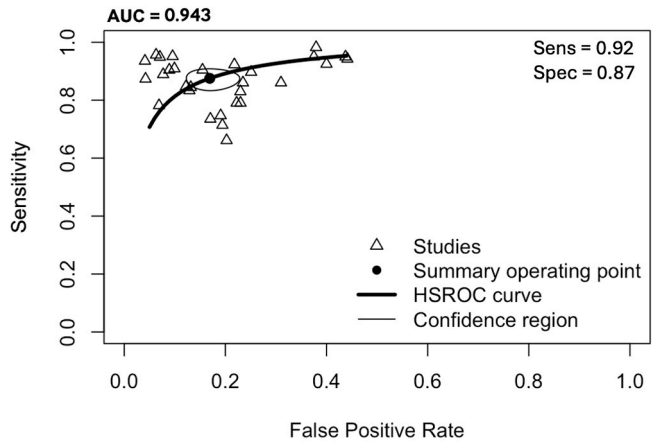

Fig. 11. Hierarchical summary receiver operating characteristic curve of machine learning algorithms for the diagnosis of thyroid carcinoma, using cytological findings as the reference standard and tested on external datasets.

more objective and reproducible image interpretation. Future research should aim to integrate AI into multimodal clinical workflows, combining ultrasound, cytological, and molecular data, while ensuring transparency, explainability, and equity across diverse patient populations.

Finally, some limitations of the present study should be acknowledged. First, all included studies were conducted in Asia. Although based on a large patient population, the findings may not be representative of the general population due to differences in genetic background, dietary habits, and iodine intake. Second, a substantial proportion of studies relied on public datasets for external validation. These open-source datasets often lack comprehensive clinical and, in particular, imaging information, potentially reducing the robustness and reliability of the results. In addition, most included studies did not report whether follow-up imaging was used as an adjunct reference standard for low-risk nodules (TIRADS 1–3). Consequently, the datasets likely over-represent nodules that underwent cytology or post-surgical histopathological evaluation (typically TIRADS 4–5), while under-representing nodules managed through surveillance. This limitation may reduce the representativeness of the included cohorts and should be taken into account when interpreting diagnostic accuracy. Finally, the underrepresentation of pathological subtypes—with most studies focusing on classical papillary thyroid cancer—has led to diagnostic inequity for patients with follicular thyroid cancer and other rare variants. This limitation may impair the generalizability of algorithmic performance across institutions, imaging devices, and multiethnic cohorts. We must also underline that an additional major limitation of our meta-analysis is the inconsistent reporting of key methodological variables across the included studies. In particular, information on TIRADS score distribution, the use of machine learning versus deep learning approaches, the specific deep learning architectures adopted, whether nodule segmentation was performed manually or automatically, and the histologic subtype was rarely provided in a standardized manner. The absence of these data prevented more detailed subgroup analyses and meta-regressions that might have explained part of the heterogeneity observed in the pooled accuracy estimates

Finally, the 'black-box' nature of AI models remains a major obstacle to clinical implementation. Current systems still fall short of providing transparent insights into their decision-making process. In particular, they fail to adequately explain the relative importance of key morphological features, such as microcalcifications versus vascular patterns [57–61]. This lack of interpretability hampers the clinical validation of AI outputs, especially in cases of misclassification—for example, the erroneous identification of Hashimoto's thyroiditis as thyroid malignancy.

In conclusion, this meta-analysis confirms that AI-based ultrasound models achieve high diagnostic accuracy in differentiating benign from malignant thyroid nodules, with performance metrics consistently validated across independent datasets. Their consistent performance across validation frameworks supports their potential integration into

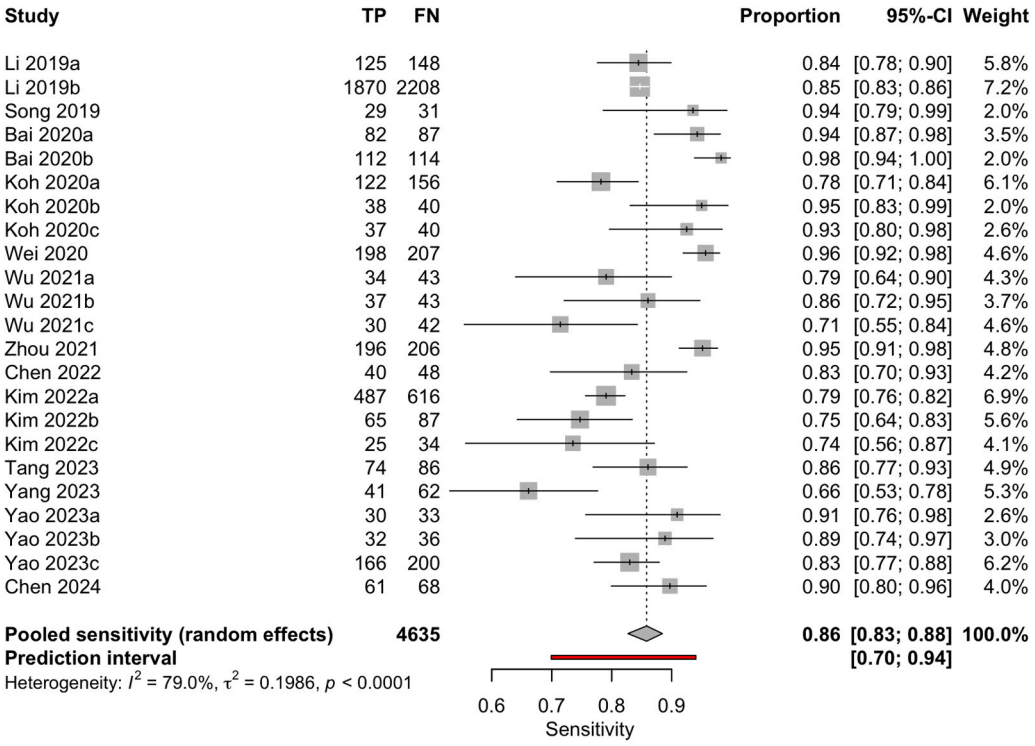

Fig. 12. Forest plot of sensitivity estimates from studies evaluating machine learning algorithms for the diagnosis of thyroid carcinoma, using histological findings as the reference standard and tested on external datasets.

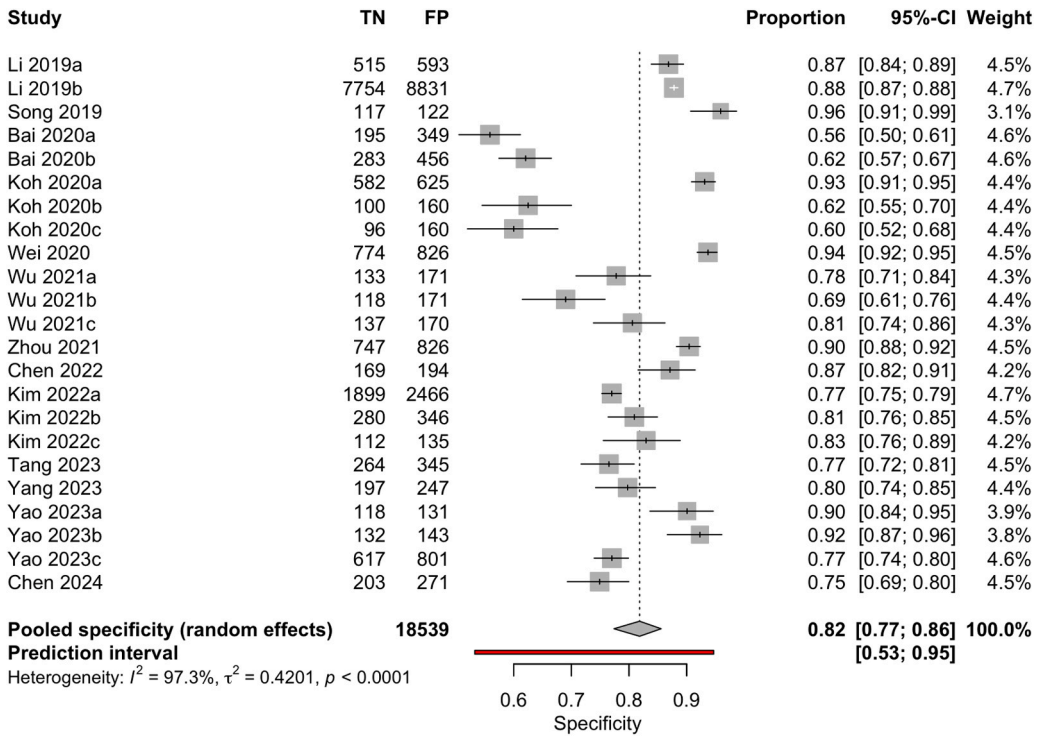

Fig. 13. Forest plot of specificity estimates from studies evaluating machine learning algorithms for the diagnosis of thyroid carcinoma, using histological findings as the reference standard and tested on external datasets.

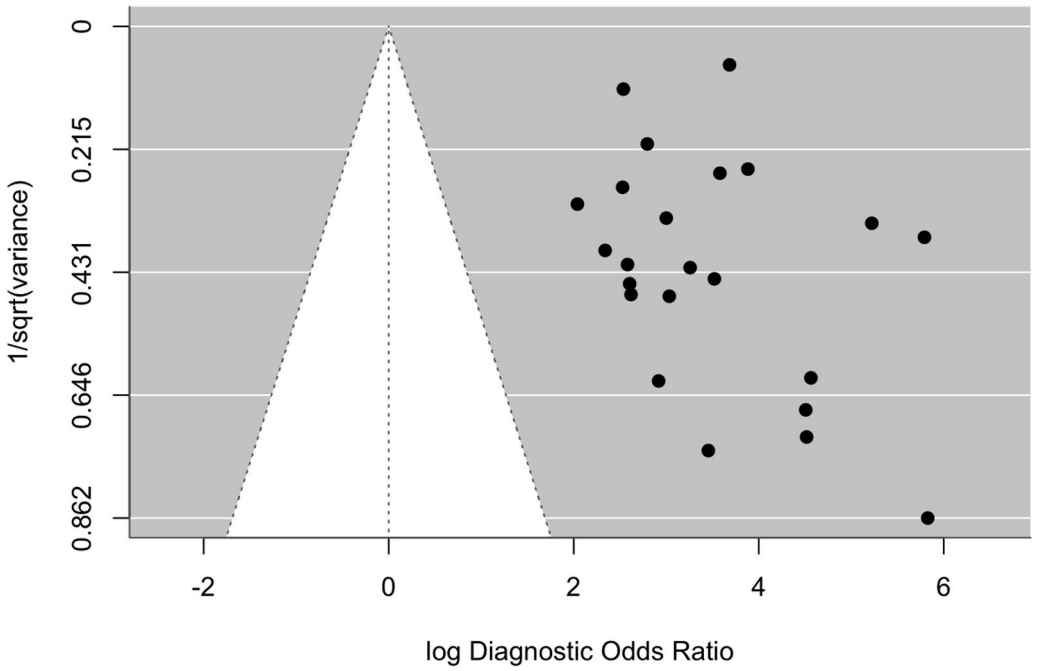

Fig. 14. Deeks' funnel plot for publication bias of studies using histological findings as the reference standard and tested on external datasets.

routine clinical workflows, paving the way for a more objective, reproducible, and data-driven approach to thyroid nodule management.

CRediT authorship contribution statement

Claudio Casella: Visualization, Validation, Supervision. **Carlo Cappelli:** Visualization, Validation, Supervision, Conceptualization. **Roberto Gatta:** Writing – original draft, Data curation. **Elisa Gatta:** Writing – original draft, Formal analysis, Data curation. **Samuele Isoli:** Writing – original draft, Data curation. **Riccardo Morandi:** Writing – review & editing. **Simone Vetrugno:** Writing – review & editing. **Sara Corvaglia:** Writing – review & editing. **Ilenia Pirola:** Visualization,

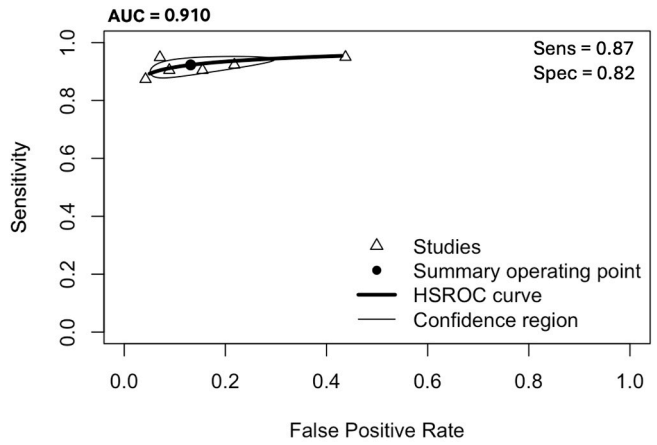

Fig. 15. Hierarchical summary receiver operating characteristic curve of machine learning algorithms for the diagnosis of thyroid carcinoma, using histological findings as the reference standard and tested on external datasets.

Validation, Supervision. **Virginia Maltese:** Writing – review & editing.

Ethical statement

As this work is a systematic review and meta-analysis based on previously published studies, it did not involve direct patient recruitment or the use of individual patient data. Therefore, approval from an institutional ethics committee was not required.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. Zhang, X.N. Wang, L. Jiang, C.X. Yu, Y.N. Chen, X.J. Yu, M.F. Pan, Conventional ultrasonography and elastosonography in diagnosis of malignant thyroid nodules: A systematic review and meta-analysis, Front Endocrinol. (Lausanne) 13 (2022) 1082881, https://doi.org/10.3389/fendo.2022.1082881.

[2] C. Cappelli, M. Castellano, I. Pirola, D. Cumetti, B. Agosti, E. Gandossi, E. Agabiti Rosei, The predictive value of ultrasound findings in the management of thyroid nodules, Qjm 100 (1) (2007) 29–35, https://doi.org/10.1093/qjmed/hcl121.

[3] Z. Wang, S.S. Huang, Y.F. Zhu, D.D. Hao, Y.Z. Zhang, C.Q. Chen, Y.W. Wang, Z. H. Jiang, F.S. Pan, J.Y. Liang, X.Y. Xie, Z. Yang, B. Li, H.P. Xiao, A new thyroid imaging reporting and data system for nodules: based on grayscale and color Doppler ultrasonography, Eur. J. Radio. 183 (2025) 111866, https://doi.org/10.1016/j.ejrad.2024.111866.

[4] N. Angelopoulos, D.G. Goulis, I. Chrisogonidis, S. Livadas, R.D. Paparodis, I. Androulakis, J.C. Jaume, I. Iakovou, The additive value of real-time elastography to thyroid ultrasound in detecting papillary carcinoma in nodules over 20 mm in diameter, Endocrine 89 (1) (2025) 177–185, https://doi.org/10.1007/s12020-025-04248-1.

[5] T. Rago, M. Scutari, V. Loiacono, F. Santini, M. Tonacchera, L. Torregrossa, R. Giannini, N. Borrelli, A. Proietti, F. Basolo, P. Miccoli, P. Piaggi, F. Latrofa, P. Vitti, Low Elasticity of Thyroid Nodules on Ultrasound Elastography Is Correlated with Malignancy, Degree of Fibrosis, and High Expression of Galectin-3 and Fibronectin-1, Thyroid 27 (1) (2017) 103–110, https://doi.org/10.1089/thy.2016.0341.

[6] C. Cappelli, I. Pirola, E. Gandossi, B. Agosti, E. Cimino, C. Casella, A. Formenti, M. Castellano, Real-time elastography: a useful tool for predicting malignancy in thyroid nodules with nondiagnostic cytologic findings, J. Ultrasound Med 31 (11) (2012) 1777–1782, https://doi.org/10.7863/jum.2012.31.11.1777.

[7] B.R. Haugen, E.K. Alexander, K.C. Bible, G.M. Doherty, S.J. Mandel, Y.E. Nikiforov, F. Pacini, G.W. Randolph, A.M. Sawka, M. Schlumberger, K.G. Schuff, S.I. Sherman, J.A. Sosa, D.L. Steward, R.M. Tuttle, L. Wartofsky, 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American thyroid association guidelines task force on thyroid nodules and differentiated thyroid cancer, Thyroid 26 (1) (2016) 1–133, https://doi.org/10.1089/thy.2015.0020.

[8] G. Russ, S.J. Bonnema, M.F. Erdogan, C. Durante, R. Ngu, L. Leenhardt, European Thyroid association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: The EU-TIRADS, Eur. Thyroid J. 6 (5) (2017) 225–237, https://doi.org/10.1159/000478927.

[9] J. de Carlos, J. Garcia, F.J. Basterra, J.J. Pineda, M. Dolores Ollero, M. Toni, P. Munarriz, E. Anda, Interobserver variability in thyroid ultrasound, Endocrine 85 (2) (2024) 730–736, https://doi.org/10.1007/s12020-024-03731-5.

[10] F.N. Tessler, W.D. Middleton, E.G. Grant, J.K. Hoang, L.L. Berland, S.A. Teefey, J. J. Cronan, M.D. Beland, T.S. Desser, M.C. Frates, L.W. Hammers, U.M. Hamper, J. E. Langer, C.C. Reading, L.M. Scoutt, A.T. Stavros, ACR Thyroid Imaging, Reporting and Data System (TI-RADS): white paper of the ACR TI-RADS Committee, J. Am. Coll. Radio. 14 (5) (2017) 587–595, https://doi.org/10.1016/j.jacr.2017.01.046.

[11] J.H. Shin, J.H. Baek, J. Chung, E.J. Ha, J.H. Kim, Y.H. Lee, H.K. Lim, W.J. Moon, D. G. Na, J.S. Park, Y.J. Choi, S.Y. Hahn, S.J. Jeon, S.L. Jung, D.W. Kim, E.K. Kim, J. Y. Kwak, C.Y. Lee, H.J. Lee, J.H. Lee, J.H. Lee, K.H. Lee, S.W. Park, J.Y. Sung, Ultrasonography diagnosis and imaging-based management of thyroid nodules: revised korean society of thyroid radiology consensus statement and recommendations, Korean J. Radio. 17 (3) (2016) 370–395, https://doi.org/10.3348/kjr.2016.17.3.370.

[12] L. Yang, C. Li, Z. Chen, S. He, Z. Wang, J. Liu, Diagnostic efficiency among Eu-/C-/ACR-TIRADS and S-Detect for thyroid nodules: a systematic review and network meta-analysis, Front Endocrinol. (Lausanne) 14 (2023) 1227339, https://doi.org/10.3389/fendo.2023.1227339.

[13] D. Toro-Tobon, R. Loor-Torres, M. Duran, J.W. Fan, N. Singh Ospina, Y. Wu, J. P. Brito, Artificial intelligence in thyroidology: a narrative review of the current applications, associated challenges, and future directions, Thyroid 33 (8) (2023) 903–917, https://doi.org/10.1089/thy.2023.0132.

[14] V. Kaul, S. Enslin, S.A. Gross, History of artificial intelligence in medicine, Gastrointest. Endosc. 92 (4) (2020) 807–812, https://doi.org/10.1016/j.gie.2020.06.040.

[15] P. Karakitsos, B. Cochand-Priollet, P.J. Guillausseau, A. Pouliakis, Potential of the back propagation neural network in the morphologic examination of thyroid lesions, Anal. Quant. Cytol. Histol. 18 (6) (1996) 494–500.

[16] R.E. Bolinger, K.J. Hopfensperger, D.F. Preston, Application of a virtual neurode in a model thyroid diagnostic network, Proc. Annu Symp. Comput. Appl. Med Care (1991) 310–314.

[17] J. Forsström, P. Nuutila, K. Irjala, Using the ID3 algorithm to find discrepant diagnoses from laboratory databases of thyroid patients, Med Decis. Mak. 11 (3) (1991) 171–175, https://doi.org/10.1177/0272989x9101100305.

[18] X. Gao, X. Ran, W. Ding, The progress of radiomics in thyroid nodules, Front Oncol. 13 (2023) 1109319, https://doi.org/10.3389/fonc.2023.1109319.

[19] V.Y. Park, E. Lee, H.S. Lee, H.J. Kim, J. Yoon, J. Son, K. Song, H.J. Moon, J. H. Yoon, G.R. Kim, J.Y. Kwak, Combining radiomics with ultrasound-based risk stratification systems for thyroid nodules: an approach for improving performance, Eur. Radio. 31 (4) (2021) 2405–2413, https://doi.org/10.1007/s00330-020-07365-9.

[20] F. Dondi, R. Gatta, G. Treglia, A. Piccardo, D. Albano, L. Camoni, E. Gatta, M. Cavadini, C. Cappelli, F. Bertagna, Application of radiomics and machine learning to thyroid diseases in nuclear medicine: a systematic review, Rev. Endocr. Metab. Disord. 25 (1) (2024) 175–186, https://doi.org/10.1007/s11154-023-09822-4.

[21] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting, D. Moher, The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, Bmj 372 (2021) n71, https://doi.org/10.1136/bmj.n71.

[22] P.F. Whiting, A.W. Rutjes, M.E. Westwood, S. Mallett, J.J. Deeks, J.B. Reitsma, M. M. Leeflang, J.A. Sterne, P.M. Bossuyt, QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies, Ann. Intern Med 155 (8) (2011) 529–536, https://doi.org/10.7326/0003-4819-155-8-201110180-00009.

[23] B. Kocak, T. Akinci D'Antonoli, N. Mercaldo, A. Alberich-Bayarri, B. Baessler, I. Ambrosini, A.E. Andreychenko, S. Bakas, R.G.H. Beets-Tan, K. Bressem, I. Buvat, R. Cannella, L.A. Cappellini, A.U. Cavallo, L.L. Chepelev, L.C.H. Chu, A. Demircioglu, N.M. deSouza, M. Dietzel, S.C. Fanni, A. Fedorov, L.S. Fournier, V. Giannini, R. Girometti, K.B.W. Groot Lipman, G. Kalarakis, B.S. Kelly, M. E. Klontzas, D.M. Koh, E. Kotter, H.Y. Lee, M. Maas, L. Marti-Bonmati, H. Müller, N. Obuchowski, F. Orlhac, N. Papanikolaou, E. Petrash, E. Pfaehler, D. Pinto Dos Santos, A. Ponsiglione, S. Sabater, F. Sardanelli, P. Seeböck, N.M. Sijtsema, A. Stanzione, A. Traverso, L. Ugga, M. Vallières, L.V. van Dijk, J.J.M. van Griethuysen, R.W. van Hamersvelt, P. van Ooijen, F. Vernuccio, A. Wang, S. Williams, J. Witowski, Z. Zhang, A. Zwanenburg, R. Cuocolo, METhodological RadiomICs Score (METRICS): a quality scoring tool for radiomics research endorsed by EuSoMII, Insights Imaging 15 (1) (2024) 8, https://doi.org/10.1186/s13244-023-01572-w.

[24] X. Li, S. Zhang, Q. Zhang, X. Wei, Y. Pan, J. Zhao, X. Xin, C. Qin, X. Wang, J. Li, F. Yang, Y. Zhao, M. Yang, Q. Wang, Z. Zheng, X. Zheng, X. Yang, C.T. Whitlow, M. N. Gurcan, L. Zhang, X. Wang, B.C. Pasche, M. Gao, W. Zhang, K. Chen, Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study, Lancet Oncol. 20 (2) (2019) 193–201, https://doi.org/10.1016/s1470-2045(18)30762-9.

[25] J. Song, Y.J. Chai, H. Masuoka, S.W. Park, S.J. Kim, J.Y. Choi, H.J. Kong, K.E. Lee, J. Lee, N. Kwak, K.H. Yi, A. Miyauchi, Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules, Med. (Baltim.) 98 (15) (2019) e15133, https://doi.org/10.1097/md.0000000000015133.

[26] W. Song, S. Li, J. Liu, H. Qin, B. Zhang, S. Zhang, A. Hao, Multitask cascade convolution neural networks for automatic thyroid nodule detection and recognition, IEEE J. Biomed. Health Inf. 23 (3) (2019) 1215–1224, https://doi.org/10.1109/jbhi.2018.2852718.
[27] Z. Bai, L. Chang, R. Yu, X. Li, X. Wei, M. Yu, Z. Liu, J. Gao, J. Zhu, Y. Zhang, S. Wang, Z. Zhang, Thyroid nodules risk stratification through deep learning based on ultrasound images, Med Phys. 47 (12) (2020) 6355–6365, https://doi.org/10.1002/mp.14543.
[28] J. Koh, E. Lee, K. Han, E.K. Kim, E.J. Son, Y.M. Sohn, M. Seo, M.R. Kwon, J. H. Yoon, J.H. Lee, Y.M. Park, S. Kim, J.H. Shin, J.Y. Kwak, Diagnosis of thyroid nodules on ultrasonography by a deep convolutional neural network, Sci. Rep. 10 (1) (2020) 15245, https://doi.org/10.1038/s41598-020-72270-6.
[29] X. Wei, M. Gao, R. Yu, Z. Liu, Q. Gu, X. Liu, Z. Zheng, X. Zheng, J. Zhu, S. Zhang, Ensemble deep learning model for multicenter classification of thyroid nodules on ultrasound images, Med Sci. Monit. 26 (2020) e926096, https://doi.org/10.12659/msm.926096.
[30] H. Zhou, Y. Jin, L. Dai, M. Zhang, Y. Qiu, K. Wang, J. Tian, J. Zheng, Differential diagnosis of benign and malignant thyroid nodules using deep learning radiomics of thyroid ultrasound images, Eur. J. Radio. 127 (2020) 108992, https://doi.org/10.1016/j.ejrad.2020.108992.
[31] S. Peng, Y. Liu, W. Lv, L. Liu, Q. Zhou, H. Yang, J. Ren, G. Liu, X. Wang, X. Zhang, Q. Du, F. Nie, G. Huang, Y. Guo, J. Li, J. Liang, H. Hu, H. Xiao, Z. Liu, F. Lai, Q. Zheng, H. Wang, Y. Li, E.K. Alexander, W. Wang, H. Xiao, Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study, Lancet Digit Health 3 (4) (2021) e250–e259, https://doi.org/10.1016/s2589-7500(21)00041-8.
[32] G.G. Wu, W.Z. Lv, R. Yin, J.W. Xu, Y.J. Yan, R.X. Chen, J.Y. Wang, B. Zhang, X. W. Cui, C.F. Dietrich, Deep Learning Based on ACR TI-RADS can improve the differential diagnosis of thyroid nodules, Front Oncol. 11 (2021) 575166, https://doi.org/10.3389/fonc.2021.575166.
[33] J. Zhu, S. Zhang, R. Yu, Z. Liu, H. Gao, B. Yue, X. Liu, X. Zheng, M. Gao, X. Wei, An efficient deep convolutional neural network model for visual localization and automatic diagnosis of thyroid nodules on ultrasound images, Quant. Imaging Med Surg. 11 (4) (2021) 1368–1380, https://doi.org/10.21037/qims-20-538.
[34] Y. Chen, Z. Gao, Y. He, W. Mai, J. Li, M. Zhou, S. Li, W. Yi, S. Wu, T. Bai, N. Zhang, W. Zeng, Y. Lu, H. Liu, An Artificial Intelligence Model Based on ACR TI-RADS Characteristics for US Diagnosis of Thyroid Nodules, Radiology 303 (3) (2022) 613–619, https://doi.org/10.1148/radiol.211455.
[35] P. Deng, X. Han, X. Wei, L. Chang, Automatic classification of thyroid nodules in ultrasound images using a multi-task attention network guided by clinical knowledge, Comput. Biol. Med 150 (2022) 106172, https://doi.org/10.1016/j.compbiomed.2022.106172.
[36] X. Han, L. Chang, K. Song, L. Cheng, M. Li, X. Wei, Multitask network for thyroid nodule diagnosis based on TI-RADS, Med Phys. 49 (8) (2022) 5064–5080, https://doi.org/10.1002/mp.15724.
[37] X.M. Keutgen, H. Li, K. Memeh, J. Conn Busch, J. Williams, L. Lan, D. Sarne, B. Finnerty, P. Angelos, T.J. Fahey, 3rd, M.L. Giger, A machine-learning algorithm for distinguishing malignant from benign indeterminate thyroid nodules using ultrasound radiomic features, J. Med Imaging 9 (3) (2022) 034501, https://doi.org/10.1117/1.Jmi.9.3.034501.
[38] Y.J. Kim, Y. Choi, S.J. Hur, K.S. Park, H.J. Kim, M. Seo, M.K. Lee, S.L. Jung, C. K. Jung, Deep convolutional neural network for classification of thyroid nodules on ultrasound: Comparison of the diagnostic performance with that of radiologists, Eur. J. Radio. 152 (2022) 110335, https://doi.org/10.1016/j.ejrad.2022.110335.
[39] X. Zhang, V.C.S. Lee, J. Rong, F. Liu, H. Kong, Multi-channel convolutional neural network architectures for thyroid cancer detection, PLoS One 17 (1) (2022) e0262128, https://doi.org/10.1371/journal.pone.0262128.
[40] C. Chen, Y. Liu, J. Yao, K. Wang, M. Zhang, F. Shi, Y. Tian, L. Gao, Y. Ying, Q. Pan, H. Wang, J. Wu, X. Qi, Y. Wang, D. Xu, Deep learning approaches for differentiating thyroid nodules with calcification: a two-center study, BMC Cancer 23 (1) (2023) 1139, https://doi.org/10.1186/s12885-023-11456-3.
[41] Z. Gao, Y. Chen, P. Sun, H. Liu, Y. Lu, Clinical knowledge embedded method based on multi-task learning for thyroid nodule classification with ultrasound images, Phys. Med Biol. 68 (4) (2023), https://doi.org/10.1088/1361-6560/acb481.
[42] L. Tang, C. Tian, H. Yang, Z. Cui, Y. Hui, K. Xu, D. Shen, TS-DSANN: Texture and shape focused dual-stream attention neural network for benign-malignant diagnosis of thyroid nodules in ultrasound images, Med Image Anal. 89 (2023) 102905, https://doi.org/10.1016/j.media.2023.102905.
[43] W. Xu, X. Jia, Z. Mei, X. Gu, Y. Lu, C.C. Fu, R. Zhang, Y. Gu, X. Chen, X. Luo, N. Li, B. Bai, Q. Li, J. Yan, H. Zhai, L. Guan, B. Gong, K. Zhao, Q. Fang, C. He, W. Zhan, T. Luo, H. Zhang, Y. Dong, J. Zhou, Generalizability and Diagnostic Performance of AI Models for Thyroid US, Radiology 307 (5) (2023) e221157, https://doi.org/10.1148/radiol.221157.
[44] Z. Yang, S. Yao, Y. Heng, P. Shen, T. Lv, S. Feng, L. Tao, W. Zhang, W. Qiu, H. Lu, W. Cai, Automated diagnosis and management of follicular thyroid nodules based on the devised small-dataset interpretable foreground optimization network deep learning: a multicenter diagnostic study, Int J. Surg. 109 (9) (2023) 2732–2741, https://doi.org/10.1097/js9.0000000000000506.
[45] J. Yao, Y. Zhang, J. Shen, Z. Lei, J. Xiong, B. Feng, X. Li, W. Li, D. Ou, Y. Lu, N. Feng, M. Yan, J. Chen, L. Chen, C. Yang, L. Wang, K. Wang, J. Zhou, P. Liang, D. Xu, AI diagnosis of Bethesda category IV thyroid nodules, iScience 26 (11) (2023) 108114, https://doi.org/10.1016/j.isci.2023.108114.
[46] S. Yao, P. Shen, T. Dai, F. Dai, Y. Wang, W. Zhang, H. Lu, Human understandable thyroid ultrasound imaging AI report system - A bridge between AI and clinicians, iScience 26 (4) (2023) 106530, https://doi.org/10.1016/j.isci.2023.106530.
[47] C. Chen, Y. Jiang, J. Yao, M. Lai, Y. Liu, X. Jiang, D. Ou, B. Feng, L. Zhou, J. Xu, L. Wu, Y. Zhou, W. Yue, F. Dong, D. Xu, Deep learning to assist composition classification and thyroid solid nodule diagnosis: a multicenter diagnostic study, Eur. Radio. 34 (4) (2024) 2323–2333, https://doi.org/10.1007/s00330-023-10269-z.
[48] N. Feng, S. Zhao, K. Wang, P. Chen, Y. Wang, Y. Gao, Z. Wang, Y. Lu, C. Chen, J. Yao, Z. Lei, D. Xu, Deep learning model for diagnosis of thyroid nodules with size less than 1 cm: a multicenter, retrospective study, Eur. J. Radio. Open 13 (2024) 100609, https://doi.org/10.1016/j.ejro.2024.100609.
[49] L. Wu, Y. Zhou, M. Liu, S. Huang, Y. Su, X. Lai, S. Bai, K. Yang, Y. Jiang, C. Cui, S. Shi, J. Xu, N. Xu, F. Dong, Video-based AI module with raw-scale and ROI-scale information for thyroid nodule diagnosis, Heliyon 10 (19) (2024) e37924, https://doi.org/10.1016/j.heliyon.2024.e37924.
[50] T. Zhou, L. Xu, J. Shi, Y. Zhang, X. Lin, Y. Wang, T. Hu, R. Xu, L. Xie, L. Sun, D. Li, W. Zhang, C. Chen, W. Wang, C. Xu, F. Kong, Y. Xun, L. Yu, S. Zhang, J. Ding, F. Wu, T. Tang, S. Zhan, J. Zhang, G. Wu, H. Zheng, D. Kong, D. Luo, US of thyroid nodules: can AI-assisted diagnostic system compete with fine needle aspiration? Eur. Radio. 34 (2) (2024) 1324–1333, https://doi.org/10.1007/s00330-023-10132-1.
[51] Y.J. Choi, J.H. Baek, H.S. Park, W.H. Shim, T.Y. Kim, Y.K. Shong, J.H. Lee, A Computer-Aided Diagnosis System Using Artificial Intelligence for the Diagnosis and Characterization of Thyroid Nodules on Ultrasound: Initial Clinical Assessment, Thyroid 27 (4) (2017) 546–552, https://doi.org/10.1089/thy.2016.0372.
[52] Y. Li, Y. Liu, J. Xiao, L. Yan, Z. Yang, X. Li, M. Zhang, Y. Luo, Clinical value of artificial intelligence in thyroid ultrasound: a prospective study from the real world, Eur. Radio. 33 (7) (2023) 4513–4523, https://doi.org/10.1007/s00330-022-09378-y.
[53] K.Z. Swan, J. Thomas, V.E. Nielsen, M.L. Jespersen, S.J. Bonnema, External validation of AIBx, an artificial intelligence model for risk stratification, in thyroid nodules, Eur. Thyroid J. 11 (2) (2022), https://doi.org/10.1530/etj-21-0129.
[54] E.J. Ha, J.H. Lee, D.H. Lee, J. Moon, H. Lee, Y.N. Kim, M. Kim, D.G. Na, J.H. Kim, Artificial Intelligence Model Assisting Thyroid Nodule Diagnosis and Management: A Multicenter Diagnostic Study, J. Clin. Endocrinol. Metab. 109 (2) (2024) 527–535, https://doi.org/10.1210/clinem/dgad503.
[55] K. Jassal, M. Edwards, A. Koohestani, W. Brown, J.W. Serpell, J.C. Lee, Beyond genomics: artificial intelligence-powered diagnostics for indeterminate thyroid nodules-a systematic review and meta-analysis, Front Endocrinol. (Lausanne) 16 (2025) 1506729, https://doi.org/10.3389/fendo.2025.1506729.
[56] C. Cappelli, I. Pirola, B. Agosti, A. Tironi, E. Gandossi, P. Incardona, F. Marini, A. Guerini, M. Castellano, Complications after fine-needle aspiration cytology: a retrospective study of 7449 consecutive thyroid nodules, Br. J. Oral. Maxillofac. Surg. 55 (3) (2017) 266–269, https://doi.org/10.1016/j.bjoms.2016.11.321.
[57] M. Reyes, R. Meier, S. Pereira, C.A. Silva, F.M. Dahlweid, H. von Tengg-Kobligk, R. M. Summers, R. Wiest, On the interpretability of artificial intelligence in radiology: challenges and opportunities, Radio. Artif. Intell. 2 (3) (2020) e190043, https://doi.org/10.1148/ryai.2020190043.
[58] Z. Salahuddin, H.C. Woodruff, A. Chatterjee, P. Lambin, Transparency of deep neural networks for medical image analysis: a review of interpretability methods, Comput. Biol. Med 140 (2022) 105111, https://doi.org/10.1016/j.compbiomed.2021.105111.
[59] F.N. Tessler, J. Thomas, Artificial intelligence for evaluation of thyroid nodules: a primer, Thyroid 33 (2) (2023) 150–158, https://doi.org/10.1089/thy.2022.0560.
[60] J.R. Geis, A.P. Brady, C.C. Wu, J. Spencer, E. Ranschaert, J.L. Jaremko, S.G. Langer, A. Borondy Kitts, J. Birch, W.F. Shields, R. van den Hoven van Genderen, E. Kotter, J. Wawira Gichoya, T.S. Cook, M.B. Morgan, A. Tang, N.M. Safdar, M. Kohli, Ethics of artificial intelligence in radiology: summary of the joint European and North American Multisociety Statement, Radiology 293 (2) (2019) 436–440, https://doi.org/10.1148/radiol.2019191586.
[61] A.P. Brady, E. Neri, Artificial intelligence in radiology-ethical considerations, Diagnostics 10 (4) (2020), https://doi.org/10.3390/diagnostics10040231.