BMC Cancer

SYSTEMATIC REVIEW Open Access

Artificial intelligence performance in ultrasound-based lymph node diagnosis: a systematic review and meta-analysis

Xinyang Han[1], Jingguo Qu[1], Man-Lik Chui[1], Simon Takadiyi Gunda[1], Ziman Chen[1], Jing Qin[2], Ann Dorothy King[3], Winnie Chiu-Wing Chu[3], Jing Cai[1] and Michael Tin-Cheung Ying[1*]

Abstract

Background and objectives Accurate classification of lymphadenopathy is essential for determining the pathological nature of lymph nodes (LNs), which plays a crucial role in treatment selection. The biopsy method is invasive and carries the risk of sampling failure, while the utilization of non-invasive approaches such as ultrasound can minimize the probability of iatrogenic injury and infection. With the advancement of artificial intelligence (AI) and machine learning, the diagnostic efficiency of LNs is further enhanced. This study evaluates the performance of ultrasound-based AI applications in the classification of benign and malignant LNs.

Methods The literature research was conducted using the PubMed, EMBASE, and Cochrane Library databases as of June 2024. The quality of the included studies was evaluated using the QUADAS-2 tool. The pooled sensitivity, specificity, and diagnostic odds ratio (DOR) were calculated to assess the diagnostic efficacy of ultrasound-based AI in classifying benign and malignant LNs. Subgroup analyses were also conducted to identify potential sources of heterogeneity.

Results A total of 1,355 studies were identified and reviewed. Among these studies, 19 studies met the inclusion criteria, and 2,354 cases were included in the analysis. The pooled sensitivity, specificity, and DOR of ultrasound-based machine learning in classifying benign and malignant LNs were 0.836 (95% CI [0.805, 0.863]), 0.850 (95% CI [0.805, 0.886]), and 33.331 (95% CI [22.873, 48.57]), respectively, indicating no publication bias ($p = 0.12$). Subgroup analyses may suggest that the location of lymph nodes, validation methods, and type of primary tumor are the sources of heterogeneity.

Conclusion AI can accurately differentiate benign from malignant LNs. Given the widespread use of ultrasonography in diagnosing malignant LNs in cancer patients, there is significant potential for integrating AI-based decision support systems into clinical practice to enhance the diagnostic accuracy.

Keywords Ultrasonography, Lymph node, Machine learning, Radiomics, Computer-aided diagnosis

*Correspondence:
Michael Tin-Cheung Ying
michael.ying@polyu.edu.hk
[1]The Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong, China
[2]Centre for Smart Health and School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China
[3]Department of Imaging and Interventional Radiology, The Chinese University of Hong Kong, Hong Kong, China

Background

Lymph nodes (LNs) are small, oval-shaped structures of the lymphatic system. Their enlargement may signify infection, inflammation, or malignancy, necessitating immediate diagnosis and intervention [1]. Typically, normal LNs are non-palpable, whereas enlarged LNs with varying degrees of firmness may indicate different pathological conditions; for instance, metastatic LNs often feel firm upon palpation [2]. Lymph node metastasis occurs when cancer cells from a primary tumor invade the lymphatic system and subsequently establish themselves within the LNs, signaling the tumor's progression and increasing aggressiveness, leading to a poorer patient prognosis [3]. Lymphoma, moreover, ranks as the second most prevalent malignancy in the head and neck region [4]. Consequently, precise and early identification of malignant LNs is crucial for accurate disease staging, effective treatment planning, and prognosis prediction [5].

Ultrasonography is a tool for assessing LNs due to its non-invasive nature, high availability, and cost-effectiveness while the sonographic characteristics of normal and abnormal LNs are not always distinct [5, 6]. When integrated with fine-needle aspiration cytology (FNAC), ultrasound can confirm LN metastasis with an accuracy rate exceeding 94% [5]. Despite its precision, FNAC remains invasive, time-intensive, and expensive, with an associated risk of infection. To minimize unnecessary biopsies and enhance diagnostic accuracy, computerized texture analysis and end-to-end ultrasound diagnostic models have been developed.

In recent years, AI and computer-aided diagnosis (CAD), which combine machine learning (ML) with medical imaging data, have shown potential for improving diagnostic efficacy. Radiomics [7] and ML techniques [8–10] are two fundamental components of CAD and have been extensively employed in the analysis of medical images, including the diagnosis of LNs in ultrasound imaging. Deep learning (DL), an advanced branch of ML, uses artificial neural networks to model and interpret complex data. Convolutional neural networks (CNNs) have been widely applied in medical image diagnosis, showing great potential for diagnosing LNs in ultrasound images by automatically extracting high-dimensional features [11]. Compared to radiomics combined with traditional ML, DL offers the advantage of learning features and making diagnoses in a more efficient, end-to-end manner.

The diagnostic accuracy of LNs in ultrasound images has been significantly improved with these AI techniques [12, 13]. However, the diagnostic accuracy of AI in the assessment of LNs in ultrasound images has not been systematically evaluated. In this study, we conducted a comprehensive pooled analysis to assess the diagnostic accuracy of ultrasound-based AI techniques for LNs, including subgroup analyses to explore study heterogeneity and enhance result interpretability. Our findings offer a valuable reference for the clinical integration of AI in LN diagnosis through ultrasound imaging.

Methods

This review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [14]. The protocol was registered on the International Prospective Register of Systematic Reviews (https://www.crd.york.ac.uk/PROSPERO/CRD42023411011).

Data sources and search strategy

We conducted a systematic literature search published between January 2010 and June 2024 using the PubMed, Embase, Medline, Google Scholar, and Cochrane Library databases. We used the following strategies to retrieve studies: "echography OR ultrasound" and "lymph node" and "texture analysis" OR "machine learning" OR "artificial intelligence" OR "computer-aided diagnosis" OR "radiomics" in titles or abstracts. In addition, references to the identified articles and review articles were searched manually for relevant studies. All references used in the included literature were also independently reviewed.

Study selection

Titles and abstracts of retrieved articles were screened for eligibility, with full texts of relevant studies reviewed and data extracted for those meeting inclusion criteria. Two reviewers (Han and Qu) independently selected the literature, resolving any differences through discussion with a third reviewer. Lists with retrieved records from the searched databases were imported into the Endnote 21 reference manager (EndNote™, Clarivate™), and duplicates were identified and removed to obtain a final list of articles.

The inclusion criteria were as follows: (1) the objective of the study was to assess the diagnostic accuracy of machine learning-based ultrasound assessment of LNs; (2) the diagnostic gold standard was pathological diagnosis; (3) the number of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) cases determined by ultrasound could be obtained from the text; (4) studies were reported in English; (5) the region of interest (ROI) is LN instead of the primary tumor.

Data extraction and sub-group analysis

The present study involved the independent extraction of data by two reviewers. The extracted data included general study details (authors, publication year, country, study design, sample size) as well as methodological

information, diagnostic test characteristics, and results, including TP, TN, FP, FN. Most studies divided their data into multiple cohorts, such as training, testing, validation, or independent external validation sets. To ensure the validity and objectivity of the assessment, only the sample sizes from the testing/validation sets and cross-validation were included. Due to the limited number of independently validated studies, internal validation data were used exclusively for this meta-analysis. When multiple ML algorithms were evaluated, only the primary proposed method demonstrating the highest area under the receiver operating characteristic curve (AUC) was included, as AUC is a robust indicator of ML performance.

In ML, considerable methodological and parameter heterogeneity across studies can introduce variability in outcomes; thus, a systematic classification was conducted to identify heterogeneity sources, with several covariates organized into six subgroups for subgroup analysis to assess their impact on outcomes and uncover unique attributes and potential heterogeneity sources. These subgroups encompassed (1) the location of the detected LNs, (2) type of disease in positive cases, (3) the utilization of AI methods (i.e., either for feature extraction or model development), (4) the method employed for ROI segmentation (i.e., manual or without segmentation), (5) the validation approach (i.e., n-fold cross-validation or random division of the validation set), (6) input for ML/DL model (i.e., radiomics/clinical features or images). For the other variables, it was decided to forgo subgroup analysis due to the large heterogeneity among the included studies. Instead, a comparative analysis was conducted by aggregating and summarizing the available information.

Quality assessment

The reviewer independently evaluated the risk of bias in each study using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) recommended by the Cochrane Collaboration [15]. The tool has four domains: "patient selection," "index test," "reference standard," and "flow and timing". Specific Signaling Questions can be found in Appendix Table S1.

Statistical analysis and software tools

This meta-analysis utilized the Open Meta-Analyst Software and StataMP 17 to analyze the data, and statistical significance was reported using 95% confidence intervals (CIs). Pooled estimates for sensitivity, specificity, diagnostic odds ratio (DOR), and positive and negative likelihood ratios (PLR and NLR), along with their corresponding 95% CIs, were employed to assess the accuracy of AI methods in detecting malignant LNs. A multivariate random-effects model was used to account for within- and between-subject variability as well as threshold effects [16]. Summary receiver operating characteristic (SROC) curves and forest plots were constructed for all analyses. Subgroup analyses were conducted to address heterogeneity, which was evaluated using the inconsistency index (I^2) and Cochran's χ^2 test (Q test). Publication bias was assessed using Deek's funnel plot [17]. TP, FP, TN, and FN values were extracted or derived from each study, and a correction factor of 0.5 was applied to zero values to address the zero-cell count issue [18].

Results

Study selection

After removing duplicates, 1,321 were excluded during screening, and 1,355 studies were identified. Several studies were excluded for the following reasons: focusing solely on a single texture feature rather than using CAD [19], using contrast-enhanced ultrasound without grayscale [20], predictive endpoints rather than the benign or malignant status of the LNs [21], employing endobronchial ultrasound [22–25], using only power Doppler sonograms [26], or lacking a machine-learning model [27]. A total of 25 studies were included in the systematic review. Five studies did not provide the complete information to calculate the confusion matrix (TP, FP, TN, and FN) [12, 28–31], and one study did not perform validation/testing [32], resulting in their exclusion from the meta-analysis. Ultimately, the remaining 19 studies were included in the quantitative analyses (meta-analysis) [13, 33–50]. Figure 1 illustrates the article screening flow chart.

Characteristics of included studies

A comprehensive summary of the extracted data can be found in Tables 1 and 2. The systematic review included 25 studies from nine countries, with only one being prospective (study design in Table 2).

Regarding the ultrasound technique in Table 2, most studies ($n = 19$) used B-mode ultrasound as the only imaging modality, while three incorporated ultrasound elastography, and three used Doppler ultrasound. For case involvement, most studies used cases without treatment or newly diagnosed, while 2 studies used post-operative cases.

As shown in Table 1, nine studies evaluated axillary LNs, while twelve and four studies examined cervical LNs and multiple body locations, respectively. Regarding patient selection, one study focused on nasopharyngeal carcinoma patients, one on non-small cell lung cancer patients, five on thyroid cancer patients, eight on breast cancer patients, and ten on patients with enlarged LNs irrespective of the primary tumor. Six studies used multicenter data, but only three of them conducted independent external validation. More specifically, only two

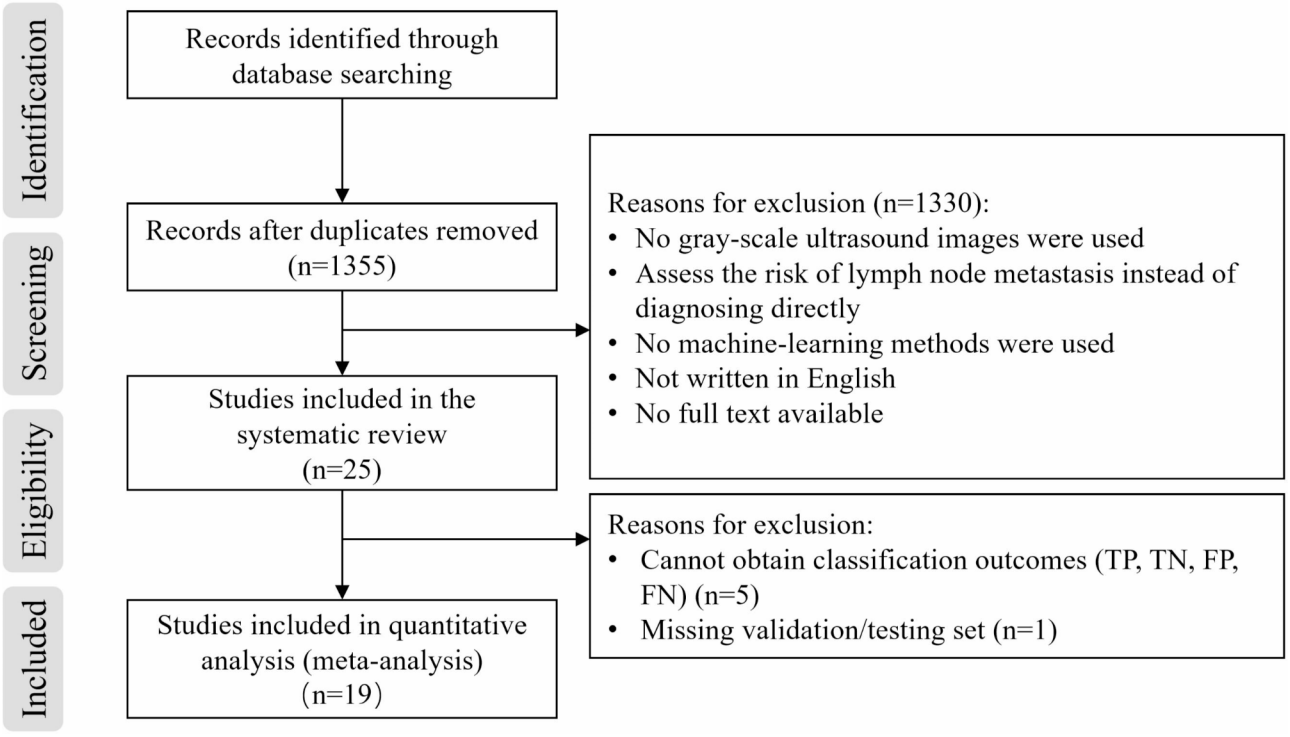


Fig. 1 Flowchart of the study selection process

studies provided classification outcomes (TP, TN, FP, FN), thus the results of external validation sets are not included in the analysis.

More Information regarding the pathological gold standard and the inclusion and exclusion criteria can be found in Appendix A and Supplementary Table S3. Overall, we included 1,152 benign and 1,202 malignant cases, with pathological biopsy as the gold standard across all studies.

Quality assessment

The quality assessment of each study is illustrated in Fig. 2. Of the 25 studies, 24 were retrospective, utilizing archived patient data, indicating a potentially high risk of selection bias. Only one study was prospective and recruited 37 patients with suspected malignant LNs [31]. As the evaluation focused on CAD accuracy with automated feature extraction and model development, and all studies used pathological results as the reference standard, the risk of bias for the index test and reference standard was considered low. One study did not allocate a separate testing set, resulting in a high risk of bias [32]. Studies involving LNs from health hy controls without pathologic confirmation, inconsistent reference standards, or excluding patients due to ongoing treatment or poor image quality were considered to have a higher risk of bias in the flow and timing domain. Specific evaluations are provided in Appendix Table S2.

Diagnostic accuracy and heterogeneity

Figure 3 presents the descriptive and pooled estimates assessing the diagnostic efficacy of AI in detecting malignant LNs using ultrasound. Of the initially included 25 studies, 19 were eligible for meta-analysis, with six excluded due to insufficient quantitative information. The meta-analysis demonstrated high diagnostic accuracy for AI methods, with pooled sensitivity of 0.836 (95% CI [0.805, 0.863]), specificity of 0.850 (95% CI [0.805,0.886]), and a diagnostic odds ratio (DOR) of 33.331 (95% CI [22.873, 48.571]). Forest plots (Fig. 3 and Appendix Figure S7) and SROC curves (Fig. 4) further support the excellent diagnostic accuracy of AI methods, while Deek's funnel plot (Fig. 5) indicated a low risk of publication bias ($p = 0.12$).

Heterogeneity was assessed using I^2 statistics and Cochran's Q test: pooled sensitivity ($Q = 29.69$, $I^2 = 39.36\%$, $p = 0.041$), specificity ($Q = 48.77$, $I^2 = 63.09\%$, $p < 0.001$), and DOR ($Q = 34.59$, $I^2 = 47.96\%$, $p = 0.011$), respectively. These results suggest moderate heterogeneity in pooled specificity and DOR and low heterogeneity in pooled sensitivity. Subgroup analyses were conducted to further investigate the sources of heterogeneity.

Subgroup analysis

Quantitative and qualitative results were reviewed for each subgroup, with subgroups containing only one study excluded from analysis (Table 3). Six aspects were considered in the subgroup analysis: location (Fig. 6A),

Table 1 Accuracy measures and subgroup information of studies for the systematic review. *Italic* studies are not included in the quantitative statistical analysis

Study	Location	Disease (Positive Cases)	Method	ROI Delineation	Center (Cases)	Validation	Train/Test	Input for ML/DL Model	Balanced Data	Benign	Malignant	Sen	Spe	TP	TN	FP	FN
Lin, 2023	Cervical	NPC	ML	Manual	1 (489)	Split*in*	7:3	Features	Y	85	61	0.887	0.698	54	59	26	7
Abbasian Ardakani, 2018	Cervical	PTC	ML	Without Segmentation	1 (340)	Split*in*	Train: 280 Test: 60	Features (Wavelet)	Y	30	30	0.933	0.966	28	29	1	2
Zhu, 2022	Cervical	Suspicious LN	DL	N/A	3 (566/105/92)	Split*ex*	Train: 395 Test*in*: 171 $Test_{ex1}$: 105 $Test_{ex2}$: 92	Images	Y	57	114	0.895	0.789	102	45	12	12
Abbasian Ardakani, 2022	Cervical	PTC	DL	Without Segmentation	2 (215/158)	Split*ex*	Train: 172 (80%) $Test_{in}$: 43 $Test_{ex1}$: 158	Images	Y	20	20	0.954	1.000	19	20	0	1
Ozaki, 2022	Axillary	Breast Cancer	DL	Without Segmentation	1 (728)	Split*in*	Train: 628 Test: 100	Images	Y	50	50	0.940	0.880	47	44	6	3
Coronado-Gutiérrez, 2022	Axillary	Breast Cancer	DL	Manual	1 (180)	Split*in*	Re-train: 71 Test: 109	Features	Y	73	36	0.778	1.000	28	73	0	8
Coronado-Gutiérrez, 2019	Axillary	Breast Cancer	DL	Manual	2 (118)	5-fold CV	-	Features	Y	65	53	0.849	0.877	45	57	8	8
Chmielewski, 2015	Axillary	Breast Cancer	ML	Manual	1 (105)	10-fold CV	-	Features	Y	81	24	0.900	0.900	22	73	8	2
Zhang, 2017	Axillary	Breast Cancer	ML	Manual	1 (161)	Leave-one-out CV	-	Features	Y	69	92	0.837	0.855	77	59	10	15
Lee, 2018	Cervical	Thyroid Cancer	DL	Without Segmentation	1 (812)	Split*in*	Train: 612 Test: 200	Features	Y	100	100	0.890	0.770	89	77	23	11
Luo, 2023	Cervical	NA	ML	Manual	1 (189)	Split*in*	7:3	Features	N	11	47	0.750	0.910	35	10	1	12
Pham, 2022	Axillary	Breast Cancer	ML	Without Segmentation	1 (107)	10-fold CV	-	Features	Y	57	50	0.840	0.947	42	54	3	8
Tahmasebi A, 2021	Axillary	Suspicious LN	DL	Without Segmentation	1 (317)	Split*in*	8:2	Images	Y	32	32	0.740	0.644	24	21	11	8
Zhang, 2008	Cervical	NA	ML	Semi-Automated Segmentation	1 (210)	10-fold CV	-	Features	Y	96	114	0.842	0.844	96	81	15	18
Chen, 2020	ALL	Suspicious LN	ML	Manual	1 (543)	Split*in*	Traini: 407 nodes Validation: 136 nodes	Features	Y	35	100	0.772	0.771	77	27	8	23
Tang, 2023	Axillary	Breast Cancer	ML	Manual	2 (130/20)	Mix and Split Validation	Traini: 105 nodes Validation: 45 nodes	Features	Y	12	33	0.727	1.000	24	12	0	9
Lyu, 2024	ALL	Suspicious LN	ML	Manual	1 (160)	Split*in*	6:4	Features	Y	23	40	0.900	0.83	36	19	4	4

Table 1 (continued)

Study	Location	Disease (Positive Cases)	Method	ROI Delineation	Center (Cases)	Validation	Train/Test	Input for ML/DL Model	Balanced Data	Benign	Malignant	Sen	Spe	TP	TN	FP	FN
Fan, 2023	Cervical	DTC (PTC: 208, FTC: 3)	ML	Manual	1 (211)	Split*in*	7:3	Features	Y	30	34	0.824	0.867	28	26	4	6
Chudobinski, 2024	ALL	Suspicious LN	DL	Without Segmentation	1 (398)	5-fold CV	-	Images	Y	226	172	0.784	0.885	135	200	26	37
Chen, 2012	Axillary & Cervical	Suspicious LN	ML	Coarse segmentation	1 (149)	5-fold CV	-	Features	Y	-	-	-	-	-	-	-	-
Drukker, 2013	Axillary	Breast Cancer	ML	Automated segmentation	1 (223)	Leave-one-out CV	-	Features	Y	-	-	-	-	-	-	-	-
Liu, 2022	Cervical	NA	ML	Manual	6 (308/38/184 /24/166/285)	Mix and Split Validation	6:4	Features	Y	-	-	-	-	-	-	-	-
Zhong, 2024	Cervical	Suspicious LN	ML	NA	1 (705)	Split*in*	Traini: 581 patients Validation: 124 patients	Features	Y	-	-	-	-	-	-	-	-
Deng, 2024	Cervical	Non-small cell lung cancer (NSCLC)	ML	Manual	3 (313 from 3 institutes)	Split*ex*	UNS	Features	N	-	-	-	-	-	-	-	-
Ardakani,2018	Cervical	PTC	ML	Automated segmentation (MaZda Software)	1 (274)	N/A	N/A	Features	Y	137	137	0.993	0.985	136	135	2	1

PTC: papillary thyroid carcinoma; DTC: differentiated thyroid carcinoma; Sen: sensitivity; Spe: specificity; CV: cross-validation; Split*in*: internal split validation; Split*ex*: external split validation; UNS: unspecific

Table 2 General characteristics of the studies included in the systematic review. *Italic* studies are not included in the quantitative statistical analysis

Study	Country	Study Design	Ultrasound Technique	DL Model	Feature Selection Methods	Feature Count	Modeling Method	Treatment
Lin, 2023	China	Retro.	B	N/A	MRMR, LASSO	7	LR	Exclude underwent treatment
Abbasian Ardakani, 2018	Iran	Retro.	B	N/A	Wavelet transform	4	SVM	Undergo total thyroidectomy, exclude irradiation, oncologic surgery
Zhu, 2022	China	Retro.	B, CD	CLA-HDM (ResNet)	N/A	N/A	N/A	Exclude history of malignancy or chemoradiation
Abbasian Ardakani, 2022	Iran	Retro.	B	ClymphNet (CNN)	N/A	N/A	N/A	Exclude surgery and radiation therapy in the neck
Ozaki, 2022	Japan	Retro.	B	Xception (GoogleNet)	N/A	N/A	N/A	Exclude hormonal therapy,chemotherapy, radiation therapy
Coronado-Gutiérrez, 2022	Spain	Retro.	B	N/A	Local binary pattern texture extraction	NA	SPLS	N/A
Coronado-Gutiérrez, 2019	Spain	Retro.	B	N/A	Fisher vector, CNN	NA	PLS	Exclude neoadjuvant chemotherapy
Chmielewski, 2015	Canada	Retro.	B	N/A	N/A	N/A	SVM	N/A
Zhang, 2017	China	Retro.	B, RTE	N/A	UNS	UNS	SVM, scoring	Exclude neoadjuvant chemotherapy
Lee, 2018	Korea	Retro.	B	CNN	N/A	N/A	N/A	Pre-operative and post-operative
Luo, 2023	China	Retro.	B	N/A	LASSO	20	LR	Exclude previous SCLN treatment (resection biopsy, radiotherapy, chemotherapy)
Pham, 2022	Malaya	Retro.	B, SWE	N/A	T-test	71	PNN	Pre-operative
Tahmasebi A, 2021	USA	Retro.	B	AutoML	N/A	N/A	N/A	N/A
Zhang, 2008	China	Retro.	B, PD	N/A	N/A	10	SVM (rough margin based)	UNS
Chen, 2020	China	Retro.	B, UE	N/A	LASSO, ANOVA	23	SVM	UNS
Tang, 2023	China	Retro.	B	N/A	ICC, LASSO	9	LASSO, Youden index	Exclude chemotherapy before ultrasound
Lyu, 2024	China	Retro.	B	N/A	Spearman test, VIF, one-way analysis of variance or Chi-square test	10	GBM	No treatment before
Fan, 2023	China	Retro.	B	N/A	ICC, Select K Best method, LASSO	13	LR, nomogram	Post-operative DTC
Chudobinski, 2024	Poland	Retro.	B	Resnte-18	N/A	N/A	N/A	N/A
Chen, 2012	China Taiwan	Pros.	B	N/A	N/A	N/A	SVM	Newly diagnosed
Drukker, 2013	USA	Retro.	B	N/A	Model performance	N/A	LDA	Newly diagnosed
Liu, 2022	China	Retro.	B	N/A	LASSO	N/A	LASSO-Vote-SVM	UNS
Zhong, 2024	China	Retro.	B, CD	BP neural network	N/A	N/A	The decision tree and BP neural network	UNS

Table 2 (continued)

Study	Country	Study Design	Ultrasound Technique	DL Model	Feature Selection Methods	Feature Count	Modeling Method	Treatment
Deng, 2024	China	Retro.	B	N/A	Univariate and multivariate analyses, ICC, Mann–Whitney U test, LASSO	15	LR, RF	N/A
Ardakani,2018	Iran	Retro.	B	N/A	Fisher, the probability of classification error, average correlation coefficients	10	1-NN	Exclude irradiation or oncological surgery

Retro: retrospective; Pros: prospective; B: B-mode; CD: colour doppler; RTE: real-time elastography; SWE: shear wave elastography; PD: power doppler; UE: ultrasound elastography; FNA: fine Needle Aspiration; CNB: core needle biopsy; SLNB: sentinel LN biopsy; CNN: convolutional neural network; UNS: unspecific; CV: cross-validation; MRMR: minimum redundancy maximum relevance; LASSO: least absolute shrinkage and selection operator; 1-NN: one-nearest neighbour; LR: logistic regression; DTC: differentiated thyroid carcinoma; SVM: support vector machine; SPLS: sparse partial least squares; PLS: partial least squares; PNN: probabilistic neural network; LDA: linear discriminant analysis; US-CNB: ultrasound-guided core needle biopsy; ICC: interclass correlation coefficient; RF: random forest; VAB: vacuum-assisted biopsy; VIF: the variance inflation factor

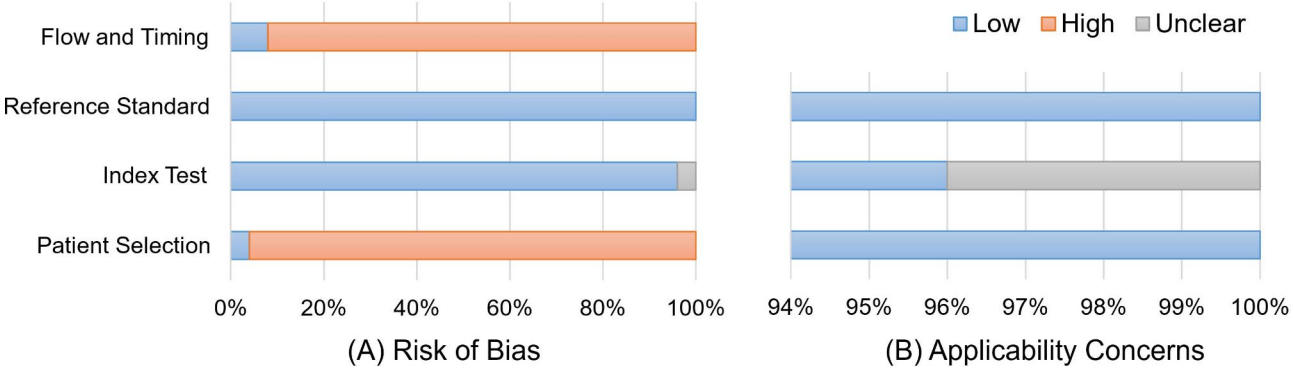


Fig. 2 Summary of the methodological quality of the included studies

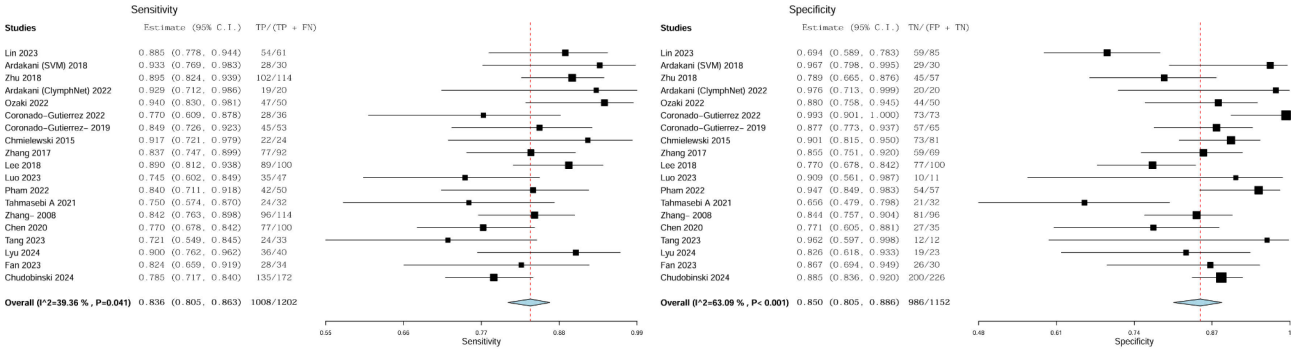


Fig. 3 Plot of individual studies sensitivity and specificity

methods (Fig. 6B), validation methods (Fig. 6C), ROI delineation approach (Fig. 6D), type of disease in positive cases (Fig. 6E), and input for the ML/DL model (Fig. 6F).

Descriptions of the six subgroup results and corresponding forest plots are provided in Appendix B. Subgroup analyses identified LN location, validation methods, and primary tumor type as sources of heterogeneity. High consistency was observed in studies with unspecified LN locations, cross-validation, and breast cancer cohorts (I^2<50%, p > 0.05). The thyroid cancer subgroup showed superior model performance with moderate heterogeneity in specificity (I^2 = 61.06%), indicating minimal variation between study results. All the summarized results of the subgroup analyses are presented in Table 3.

Discussion

This study evaluated the diagnostic efficacy of AI techniques for ultrasound imaging in LN assessment. Across 2,354 LNs from the included studies, pooled sensitivity

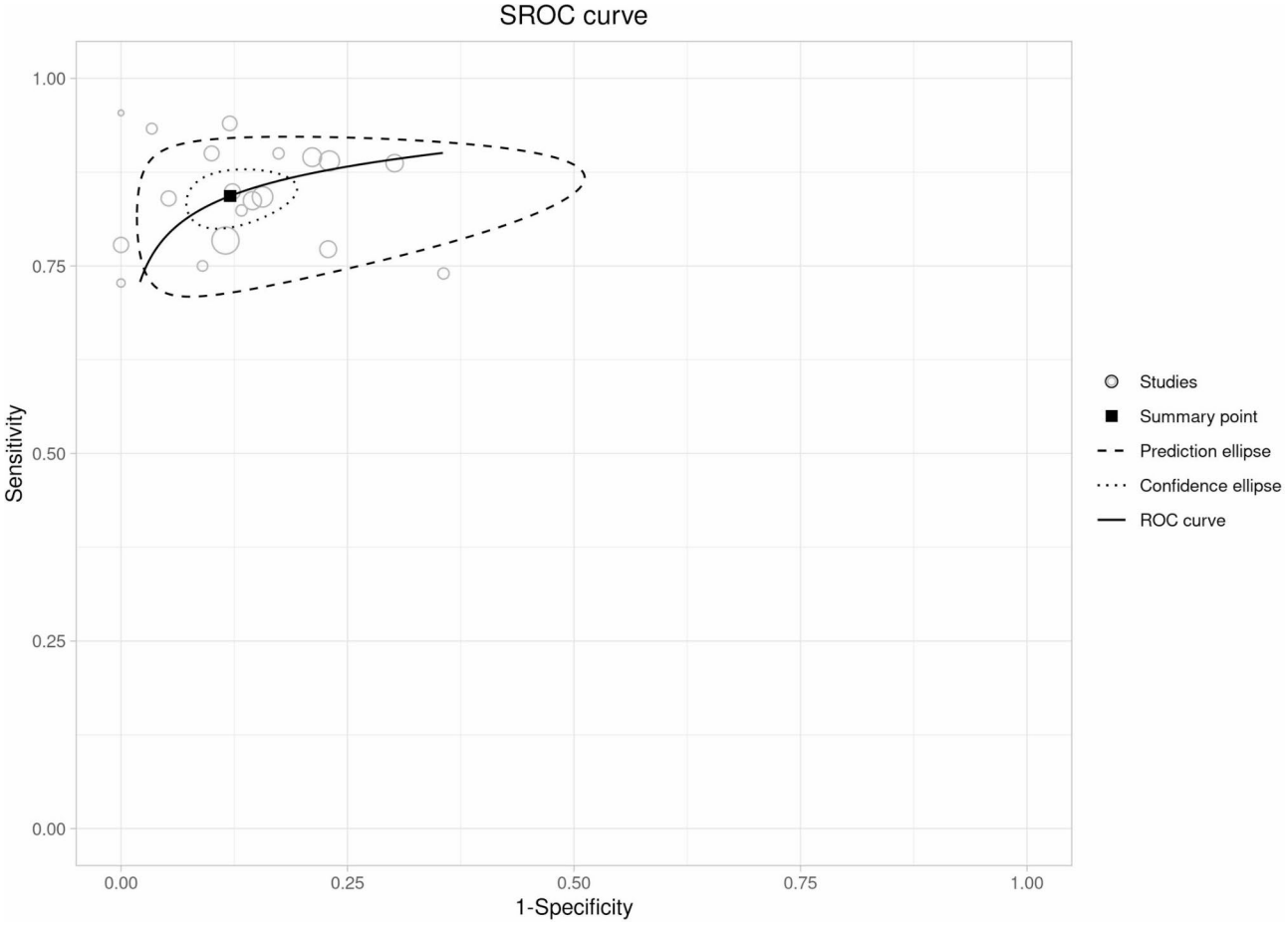


Fig. 4 Plot of summary receiver operator characteristics

and specificity were 0.836 and 0.850, respectively, for differentiating benign from malignant LNs. These findings indicate that the AI algorithms improve the accuracy of LN diagnosis in ultrasound images. To the best of our knowledge, this meta-analysis is the first study to specifically address the usage of ultrasound imaging in the CAD of nodal malignancy.

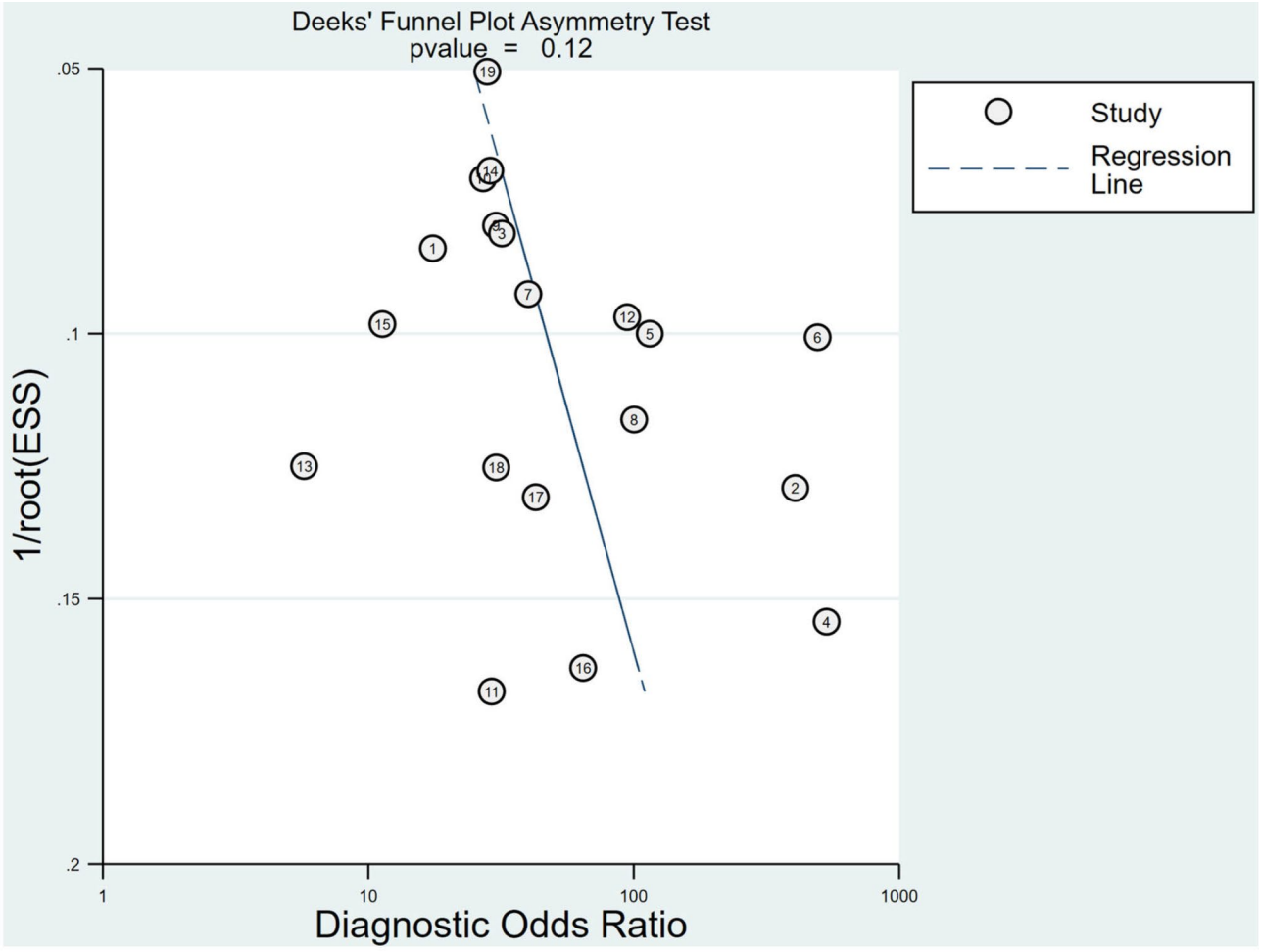
Additionally, included studies indicate different levels of heterogeneity (low for sensitivity, high for specificity). This result is anticipated given that ML and DL methods are often viewed as "black boxes" due to their complex, abstract, and data-oriented nature, which can be significantly influenced by the training data. Furthermore, there is no universally recognized procedure and hyper-parameter setting for conducting model training, although some researchers have proposed general frameworks for the research process [51].

Subgroup analysis

Subgroup analyses were conducted on various covariates to address the heterogeneity observed among the included studies. Although the primary goal in all studies was to classify cases as benign or malignant, each case may involve multiple factors, such as the patient's underlying condition, which could influence diagnostic model performance. Therefore, subgroup analyses were performed to assess whether patient-related, lesion-specific, or other factors affected the diagnostic outcomes. Our analysis indicated that the primary sources of heterogeneity stem from case selection, including LN location and primary tumor type, as well as the methods used for validation.

Lymph node location

Studies grouped by LN location exhibited lower heterogeneity, suggesting that variations across different sites may contribute to the observed heterogeneity. Most LN evaluations focused on specific anatomical sites, primarily the neck or axilla. Chen et al. [31] investigated both cervical and axillary LNs, whereas Chen et al. [48] expanded the analysis to include LNs from multiple regions, such as the axilla, neck, and groin. This focus can be attributed to the higher likelihood of nodal metastasis and the accessibility of LNs in these regions.



yb	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
Bias	12.6476	7.700382	1.64	0.119	-3.598786	28.89399
Intercept	2.582279	.7136549	3.62	0.002	1.076598	4.087959

Fig. 5 Deeks' funnel plot

Methods for modeling

The methods used for data analysis and modeling varied across studies due to different AI approaches, making it impractical to categorize each method individually. Studies were grouped by DL and traditional ML methods, with minimal differences in diagnostic performance between the two groups, indicating that model performance was not significantly affected by the choice of DL or ML.

The included studies employed various ML methods. SVM was commonly used to establish the final models [47, 49]. Some studies developed DL models, such as Abbasian Ardakani et al. [50] introducing "ClymphNet" a CNN-based model. Earlier studies often combined traditional radiological features, including echogenicity, margin, and shape, with textural features, which is a standard approach in radiomics modeling [49, 32]. Notably, Abbasian et al. [50] adopted multiple ML algorithms that were trained on quantitative and semi-quantitative (semantic) imaging features to classify LNs. These features encompass aspect ratio, prominent echogenicity, homogeneity, and microcalcification. Comparatively, their proprietary DL model outperformed the SVM protocol in internal and external validations.

Validation methods

The cross-validation subgroup showed significantly lower heterogeneity (I^2=0%, $p>0.05$) compared to the internal validation subgroup, suggesting that different validation method contributes to the observed heterogeneity.

Table 3 Summary results of subgroup analyses

Sub-type	Group	Study Count	Sen (95% CI)	Q	I^2	*P*-value	Spe (95% CI)	Q	I^2	*P*-value
Location	Cervical	8	**0.862 (0.820, 0.896)**	9.9	29.15%	0.195	0.816 (0.743, 0.871)	15.2	54.01%	0.033
	Axillary	8	0.825 (0.769, 0.869)	10.2	31.47%	0.177	**0.884 (0.808, 0.932)**	20.5	65.88%	0.005
	Unspecified	3	0.796 (0.730, 0.850)	3.0	33.48%	0.106	0.847 (0.761, 0.905)	3.6	44.20%	0.167
Method	ML	11	0.826 (0.786, 0.860)	13.5	26.04%	0.196	**0.857 (0.796, 0.903)**	24.6	59.38%	0.006
	DL	8	**0.850 (0.794, 0.893)**	15.8	55.75%	0.027	0.844 (0.766, 0.899)	24.1	70.94%	0.001
Validation Method	Split	13	0.842 (0.794, 0.881)	25.7	53.31%	0.012	0.822 (0.754, 0.874)	27.7	56.68%	0.006
	CV	6	0.823 (0.787, 0.854)	3.8	0.00%	0.576	**0.879 (0.849, 0.903)**	4.3	0.00%	0.510
ROI Delineation Approach	Manual	10	0.813 (0.771, 0.850)	11.3	20.08%	0.258	0.853 (0.782, 0.904)	23.2	61.14%	0.006
	Without Segmentation	7	**0.859 (0.793, 0.906)**	13.7	56.05%	0.034	**0.868 (0.779, 0.925)**	24.1	75.09%	< 0.001
	Unspecified	1	-	-	-	-	-	-	-	-
	Automated Segmentation	1	-	-	-	-	-	-	-	-
Type of Disease in Positive Cases	NPC	1	-	-	-	-	-	-	-	-
	Thyroid Cancer	4	**0.884 (0.828, 0.924)**	2.3	0.00%	0.510	**0.884 (0.734, 0.954)**	7.7	61.06%	0.053
	Unknown Primary Tumor	7	0.814 (0.763, 0.856)	11.6	48.46%	0.070	0.816 (0.748, 0.870)	13.2	54.58%	0.040
	Breast Cancer	7	0.834 (0.776, 0.880)	9.0	33.03%	0.176	**0.897 (0.852, 0.930)**	7.7	22.02%	0.261
Input for ML/DL model	Features	14	0.831 (0.797, 0.860)	17.3	25.03%	0.184	**0.857 (0.804, 0.898)**	33.8	61.53%	0.001
	Images	5	**0.858 (0.769, 0.917)**	12.3	67.60%	0.015	0.835 (0.725, 0.906)	14.4	72.31%	0.006

Sen: sensitivity; Spe: specificity; CV: cross-validation

Bold represents the best result in the same subgroup; *Underlined* indicates lower heterogeneity; Dark shading represents possible sources of heterogeneity

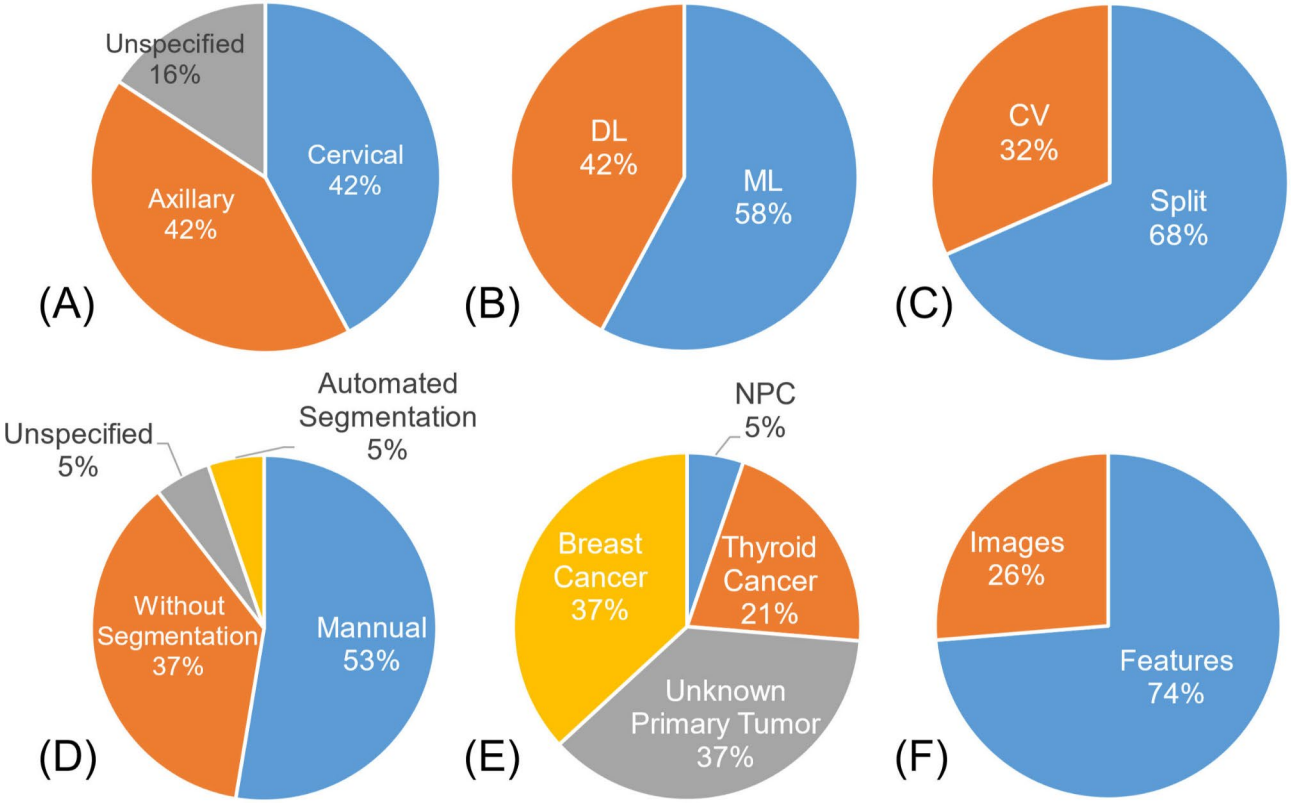


Fig. 6 A visual summary of the included studies (Quantitative assessment only). (**A**) Location, (**B**) Method, (**C**) Validation Method, (**D**) ROI Delineation Approach, (**E**) Type of Disease in Positive Cases, (**F**) Input for ML/DL model

Cross-validation, which uses each subset of data for both training and validation, reduces the risk of overfitting and enhances the model's generalizability. As a result, models validated with cross-validation tend to perform more reliably on data with different distributions. Future research should incorporate independent test sets from multiple centers to improve model credibility. Additionally, all studies used private datasets for training and validation without employing or sharing public datasets, limiting the ability to model comparison, external validation and further refinement.

ROI delineated approach

Various CAD methods, particularly ML, depend on the relevant information accurately extracted from images, which requires precise delineation of the ROIs. Different methodologies for ROI contour have been used across studies. In radiomics feature extraction, ROIs must be manually outlined by radiologists by segmenting the LNs from the background, which is a time-consuming and labor-intensive process. Despite its complexity, manual ROI delineation by experienced radiologists remains the most accurate and widely used method in CAD [12, 33, 40, 41, 44, 45, 47, 48].

To improve efficiency and reduce labor costs, some studies have explored automated and semi-automated segmentation methods. For instance, Ardakani et al.. utilized the open-source MaZda software for automated segmentation and feature extraction [32], while Drukkler et al. [29]. used a breast cancer segmentation model for automated LN segmentation. It is generally accepted that an overlap ratio of 0.4 or higher in computer-derived segmentations is sufficient for further analysis [53]. Zhang et al. [34]. deemed fully automated segmentation to be unreliable and therefore utilized a semi-automated segmentation method that involved manual coarse segmentation to generate an initial contour based on several marks. Meanwhile, Ardakani et al. [31] excluded the hyperechoic hilum of LN from the ROI for image feature extraction.

Some studies are adopting end-to-end methods, where images are directly inputted into the system, and the output provides the classification results [36–38, 42, 50]. This methodology necessitates merely a rough delineation of the ROI or without any processing, typically executed as a bounding box. All studies, except Pham et al. [37], employed convolutional neural networks (CNNs) for modeling, and utilized class activation mapping (CAM) for visualization purposes. This approach allows for the generation of final classification results without the need for delineation of ROI [37]. Our analysis suggests that diagnostic performance is better in the group that does not require manual segmentation. Thus, more effort should be put into the development of fully automated segmentation methods.

Type of disease (primary tumor type)

For the selection of LN images to be included, we recommended that LNs be divided into two categories: normal LNs which do not require pathological confirmation and exhibit clear benign LN features, and suspicious LNs which necessitate confirmation through pathological biopsy. The latter category encompasses malignant LNs (involvement of lymphoma or metastatic malignancy) and enlarged normal LNs (associated with tuberculosis or reactive LNs). Concerning the source of data on benign LNs, existing studies have not yet differentiated between normal LNs and enlarged LNs requiring biopsy, discriminating capabilities of the constructed models between these two differing types of normal LNs is questionable. However, the exploration of this aspect in current research is limited, likely due to the retrospective nature of the studies and the limitation of patient sources.

Most studies primarily focused on evaluating malignant LNs in patients diagnosed with specific types of cancer, such as breast or thyroid cancer. However, a subset of studies expanded their scope to include the assessment of unexplained cervical lymphadenopathy (CLA) [13]. Furthermore, certain studies incorporated healthy control populations to compare with the malignant LNs [44]. Notably, Coronado-Gutiérrez et al. [45] also included reactive LNs resulting from COVID-19 vaccination to highlight visual similarities between malignant nodes and LNs affected by the vaccination. Moreover, the primary tumor of three studies [40, 34, 12] is unspecified.

Regarding patient selection, the cases were discussed separately according to the status of LNs (positive or negative) instead of differentiated by the type of primary tumor. Some studies exclusively included malignant LNs from a single disease type, such as thyroid cancer, breast cancer, or nasopharyngeal cancer. Besides these studies, Ozaki et al. [38], Chen et al. [51], and Coronado et al. [45] used normal LNs from healthy individuals as the control group in their model training, while others used proven benign LNs from cancer patient [50, 49, 47, 43]. The results of the subgroup analyses suggest that variations in primary tumor type also contribute to the observed heterogeneity.

Treatment

Most studies in this meta-analysis used ultrasound images from patients before treatment for model training. Treatments such as radiotherapy and chemotherapy can alter lymphatic structures, potentially affecting LN characteristics in ultrasound images. Since ultrasonography is commonly used as a follow-up diagnostic tool in oncology, the accuracy of models trained solely on

pre-treatment images may be limited when applied to follow-up assessments, warranting further investigation. While most studies in this review excluded cases involving irradiation or oncologic surgery (Table 3), Lee et al. [42] included patients from both pre- and post-operative stages, and Fan et al. [43] only used post-operative thyroid cancer patients.

Data balance

Imbalanced data occurs when classes in a dataset are not equally represented, which can negatively affect the performance of AI models, particularly for classification tasks [54]. While various resampling techniques have been proposed to address this issue, the best approach is to avoid it during data collection. Among the included studies, only two utilized imbalanced data [30, 40], while the others implemented strategies to mitigate this problem. For instance, Ozaki et al. [38] increased the sample size by using multiple LN images from a single patient, and Ardakani et al. [32]. attempted to balance data to minimize discrepancy in the amount of cases by the non-randomization process.

Limitations

This meta-analysis and systematic review have four limitations. First, we only calculated the binary confusion matrix for benign and malignant classification in three studies that involved in multiclassification [13, 39, 48], which may lead to an overestimation of diagnostic indicators. Second, certain subgroup analyses could not be performed due to the limited number of available studies ($n \leq 3$), potentially affecting the diversity of the findings. Third, the meta-analysis primarily relied on the results yielded from internal test sets while external test sets were excluded to ensure consistency; this may have contributed to an overestimation of diagnostic accuracy. Finally, all studies were trained on private datasets, and 96% of them were retrospective, which may lead to selection bias and data loss. To enhance the robustness of CAD models, it is generally considered more reliable to use data from multicenter. Therefore, future prospective multicenter studies are needed to validate these findings.

Conclusion

In conclusion, this meta-analysis and systematic review evidence that the integration of ultrasound imaging and AI achieved high sensitivity, specificity, and accuracy in distinguishing malignant and benign LNs. There is a high potential for this integrated technique to be used in assisting clinical decisions, planning treatment strategies, and predicting the prognosis of cancer patients.

Abbreviations

AI	Artificial intelligence
LN	Lymph node
FNAC	Fine-needle aspiration cytology
CAD	Computer-aided diagnosis
ML	Machine learning
DL	Deep learning
CNN	Convolutional neural network
TP	True positive
FP	False positive
FN	False negative
TN	True negative
ROI	The region of interest
AUC	Area under the receiver operating characteristic curve
DOR	Diagnostic odds ratio

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-025-13447-y.

Supplementary Material 1

Acknowledgements

Not applicable.

Author contributions

Xinyang Han: Conceptualization, Investigation, Writing original draft, Writing review & editing. Jingguo Qu: Investigation, Writing - review & editing. Man-Lik Chui: Data curation. Simon Takadiyi Gunda: Resources. Ziman Chen: Resources. Jing Qin: Supervision. Ann Dorothy King: Supervision. Winnie Chiu-Wing Chu: Supervision. Jing Cai: Supervision. Michael Tin-Cheung Ying: Conceptualization, Supervision, Writing - review & editing.

Funding

This study has received funding by General Research Fund of the Research Grant Council of Hong Kong (Reference no. 15102222).

Data availability

This published article and its supplementary information files include all data generated or analysed during this study.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 15 November 2024 / Accepted: 3 January 2025
Published online: 13 January 2025

References

1. Chau I, et al. Rapid access multidisciplinary lymph node diagnostic clinic: analysis of 550 patients. Br J Cancer. 2003;88(3):354–61.
2. Priya NS. Lymph nodes in health and disease - A pathologist's perspective. J Oral Maxillofac Pathol. 2023;27(1):6–11.
3. Zhou H, Lei PJ, Padera TP. Progression of Metastasis through Lymphatic System. Cells, 2021. 10(3).
4. DePeña CA, Van Tassel P, Lee YY. Lymphoma of the head and neck. Radiol Clin North Am. 1990;28(4):723–43.
5. Ahuja AT, et al. Ultrasound of malignant cervical lymph nodes. Cancer Imaging. 2008;8(1):48–56.
6. Ahuja AT, Ying M. Sonographic evaluation of cervical lymph nodes. Am J Roentgenol. 2005;184(5):1691–9.
7. van Timmeren JE, et al. Radiomics in medical imaging—how-to guide and critical reflection. Insights into Imaging. 2020;11(1):91.

8. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
9. Ho TK. Random decision forests. in Proceedings of 3rd international conference on document analysis and recognition. 1995. IEEE.
10. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21–7.
11. LeCun Y, et al. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1989;1(4):541–51.
12. Liu Y, et al. Ultrasound-based Radiomics can classify the etiology of cervical lymphadenopathy: a Multi-center Retrospective Study. Front Oncol. 2022;12:856605.
13. Zhu Y, et al. Deep learning radiomics of dual-modality ultrasound images for hierarchical diagnosis of unexplained cervical lymphadenopathy. BMC Med. 2022;20(1):269.
14. Liberati A, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009;339:b2700.
15. Whiting PF, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–36.
16. Lee J, et al. Systematic review and Meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical Researchers-Part II. Statistical methods of Meta-Analysis. Korean J Radiol. 2015;16(6):1188–96.
17. Song F, et al. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. Int J Epidemiol. 2002;31(1):88–95.
18. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat Med. 2004;23(9):1351–75.
19. Kawashima Y, et al. Efficacy of texture analysis of ultrasonographic images in the differentiation of metastatic and non-metastatic cervical lymph nodes in patients with squamous cell carcinoma of the tongue. Oral Surg Oral Med Oral Pathol Oral Radiol. 2023;136(2):247–54.
20. Jiang Z, et al. Classification of superficial suspected lymph nodes: non-invasive radiomic model based on multiphase contrast-enhanced ultrasound for therapeutic options of lymphadenopathy. Quant Imaging Med Surg. 2024;14(2):1507–25.
21. Wei W, et al. Deep learning radiomics for prediction of axillary lymph node metastasis in patients with clinical stage T1-2 breast cancer. Quant Imaging Med Surg. 2023;13(8):4995–5011.
22. Nguyen P, et al. Optical differentiation between malignant and Benign Lymphadenopathy by Grey Scale Texture Analysis of Endobronchial Ultrasound Convex Probe images. Chest. 2012;141(3):709–15.
23. Ozcelik N, et al. Can artificial intelligence distinguish between malignant and benign mediastinal lymph nodes using sonographic features on EBUS images? Curr Med Res Opin. 2020;36(12):2019–24.
24. Nguyen P, et al. Optical differentiation of malignant and benign lymphadenopathy by greyscale texture analysis of endobronchial ultrasound convex probe images. Respirology. 2011;16:17.
25. Li J, et al. Deep learning with convex probe endobronchial ultrasound multimodal imaging: a validated tool for automated intrathoracic lymph nodes diagnosis. ENDOSCOPIC ULTRASOUND. 2021;10(5):361–.
26. Ying M, Cheng SC, Ahuja AT. Diagnostic accuracy of computer-aided Assessment of Intranodal Vascularity in distinguishing different causes of cervical Lymphadenopathy. Ultrasound Med Biol. 2016;42(8):2010–6.
27. Cheng SCH, Ahuja AT, Ying M. Quantification of intranodal vascularity by computer pixel-counting method enhances the accuracy of ultrasound in distinguishing metastatic and tuberculous cervical lymph nodes. Volume 9. QUANTITATIVE IMAGING IN MEDICINE AND SURGERY; 2019. pp. 1773–80. 11.
28. Zhong R, et al. Development, Validation, and comparison of 2 Ultrasound feature-guided machine learning models to Distinguish Cervical Lymphadenopathy. Ultrasound Q. 2024;40(1):39–45.
29. Drukker K, et al. Quantitative ultrasound image analysis of axillary lymph node status in breast cancer patients. Int J Comput Assist Radiol Surg. 2013;8(6):895–903.
30. Deng Z, et al. Ultrasound-based radiomics machine learning models for diagnosing cervical lymph node metastasis in patients with non-small cell lung cancer: a multicentre study. BMC Cancer. 2024;24(1):536.
31. Chen S-J, et al. Characterizing the major sonographic textural difference between metastatic and common benign lymph nodes using support vector machine with histopathologic correlation. Clin Imaging. 2012;36(4):353–e3592.
32. Ardakani AA, et al. Differentiation between metastatic and tumour-free cervical lymph nodes in patients with papillary thyroid carcinoma by grey-scale sonographic texture analysis. Pol J Radiol. 2018;83:e37–46.
33. Zhang Q, et al. Dual-modal computer-assisted evaluation of axillary lymph node metastasis in breast cancer patients on both real-time elastography and B-mode ultrasound. Eur J Radiol. 2017;95:66–74.
34. Zhang J, et al. Computer-aided diagnosis of cervical lymph nodes on ultrasonography. Comput Biol Med. 2008;38(2):234–43.
35. Tang YL, et al. Ultrasound radiomics based on axillary lymph nodes images for predicting lymph node metastasis in breast cancer. Front Oncol. 2023;13:1217309.
36. Tahmasebi A, et al. Assessment of Axillary Lymph Nodes for Metastasis on Ultrasound using Artificial Intelligence. Ultrason Imaging. 2021;43(6):329–36.
37. Pham TH et al. Fusion of B-mode and shear wave elastography ultrasound features for automated detection of axillary lymph node metastasis in breast carcinoma. EXPERT Syst, 2022. 39(5).
38. Ozaki J, et al. Deep learning method with a convolutional neural network for image classification of normal and metastatic axillary lymph nodes on breast ultrasonography. Japanese J Radiol. 2022;40(8):814–22.
39. Lyu S, et al. Application of machine-learning based on Radiomics Features in Differential diagnosis of superficial lymphadenopathy. Curr Med Imaging; 2024.
40. Luo J, et al. Clinical features combined with ultrasound-based radiomics nomogram for discrimination between benign and malignant lesions in ultrasound suspected supraclavicular lymphadenectasis. Front Oncol. 2023;13:1048205.
41. Lin M, et al. Using ultrasound radiomics analysis to diagnose cervical lymph node metastasis in patients with nasopharyngeal carcinoma. Eur Radiol. 2023;33(2):774–83.
42. Lee JH, et al. Deep learning-based computer-aided diagnosis system for localization and diagnosis of metastatic lymph nodes on Ultrasound: a pilot study. Thyroid. 2018;28(10):1332–8.
43. Fan F et al. Integration of ultrasound-based radiomics with clinical features for predicting cervical lymph node metastasis in postoperative patients with differentiated thyroid carcinoma. Endocrine, 2023.
44. Coronado-Gutiérrez D, et al. Quantitative Ultrasound Image Analysis of Axillary Lymph Nodes to diagnose metastatic involvement in breast Cancer. Ultrasound Med Biol. 2019;45(11):2932–41.
45. Coronado-Gutiérrez D, et al. Quantitative ultrasound image analysis of axillary lymph nodes to differentiate malignancy from reactive benign changes due to COVID-19 vaccination. Eur J Radiol. 2022;154:110438.
46. Chudobinski C et al. Enhancements in Radiological Detection of Metastatic Lymph Nodes Utilizing AI-Assisted Ultrasound Imaging Data and the Lymph Node Reporting and Data System Scale. Cancers (Basel), 2024. 16(8).
47. Chmielewski A, Dufort P, Scaranelo AM. A computerized system to assess Axillary Lymph Node Malignancy from Sonographic images. Ultrasound Med Biol. 2015;41(10):2690–9.
48. Chen Y, et al. Dual-mode ultrasound radiomics and intrinsic imaging phenotypes for diagnosis of lymph node lesions. Ann Transl Med. 2020;8(12):742.
49. Abbasian Ardakani A, Reiazi R, Mohammadi A. A clinical decision support system using Ultrasound textures and Radiologic features to Distinguish Metastasis from Tumor-Free Cervical Lymph nodes in patients with papillary thyroid carcinoma. J Ultrasound Med. 2018;37(11):2527–35.
50. Abbasian Ardakani A et al. Diagnosis of metastatic lymph nodes in patients with papillary thyroid Cancer: a comparative Multi-center Study of Semantic features and deep learning-based models. J Ultrasound Med, 2022.
51. Chen PC, Liu Y, Peng L. How to develop machine learning models for healthcare. Nat Mater. 2019;18(5):410–4.
52. Som PM. Detection of metastasis in cervical lymph nodes: CT and MR criteria and differential diagnosis. AJR Am J Roentgenol. 1992;158(5):961–9.
53. Horsch K, et al. Automatic segmentation of breast lesions on ultrasound. Med Phys. 2001;28(8):1652–9.
54. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Progress Artif Intell. 2016;5(4):221–32.

Publisher's note